ANALYSIS OF THE PROM ALGORITHM AS A TOOL TO GENERATE GENOME-SCALE METABOLIC-REGULATORY NETWORKS

BY

BOZENA CABALLERO

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Chemical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Adviser:

Assistant Professor Nathan D. Price

Abstract

In this paper, we analyzed the capability of PROM's algorithm to generate genome-scale metabolic-regulatory networks, which accurately predict growth phenotypes of transcriptional regulatory mutants under various conditions. *E. coli*, *M. tuberculosis* and *S. cerevisase* were used as model organisms. We showed that PROM could be successfully applied to model less complex systems (*E. coli* and *M. tuberculosis*) but not eukaryotes (*S. cereavisae*). The effects of the accuracy of the metabolic and regulatory networks reconstructions as well as the amount of gene expression data (microarrays) on PROM's ability to simulate growth phenotypes was analyzed. It was determined that well defined metabolic model and transcriptional regulatory network were crucial for PROM to be predictive. However, accurately represented gene-transcription factor (TF) interactions played a more significant role than the metabolic model. Also, those interactions had to be determined experimentally and not through an inference algorithm (such as ASTRIX). In case of the amount of gene expression data, it was observed that a number of microarrays needed for best PROM's performance was species specific and incorporation of additional samples resulted in no further improvement of the model. The extension of PROM's algorithm to predict changes in reaction rates (fluxes) for transcriptional regulatory mutants growing on different media showed that incorporation of Flux Variability Analysis (FVA) was not sufficient for such studies.

Contents

1. Modeling metabolic-regulatory networks: background and scope of this project

1.1 Introduction to systems biology

Over a decade ago, systems biology emerged as an approach to decoding life. In contrast of studying particular genes or proteins one at the time, it investigates the behavior and relationships of all of the elements in a particular biological system[1]. Through comprehensive understanding of all those interactions, systems biology aims to develop a computational model which successfully mimics cell behavior. Such achievement would have a great impact on the understanding of mechanisms underlying various diseases (e.g. cancer, tuberculosis). Specifically, it would aid in the determination of prospective drug targets, which consequently would lead to more effective therapies.

The progress made within the past decade in studying individual intracellular biochemical reaction networks was significant[2]. More than 50 genome-scale metabolic model reconstructions have been published[3]. Also, considerable efforts were made in generating transcriptional regulatory networks[4]. However, the real challenge lies in successfully integrating those two types of networks. Such genome-scale metabolic-regulatory models would allow us to better foresee how genetic mutations and transcriptional perturbations are translated into flux responses at the metabolic level[5]. There have been few successful genome-scale analysis methods published[5,6]; however, substantial progress is still required for those methods to be widely applied to less studied organisms.

1.2 Genome-scale metabolic-regulatory reconstructions

The first combined metabolic-regulatory *in silico* model was reconstructed for *Escherichia coli* using regulatory flux balance analysis (rFBA) [7]. This method is an extension of the flux-balance analysis (FBA)[8] – the most commonly used constraint based technique to model metabolic networks. The success of the FBA technique lies in restraining the number of functional states a network can take by introducing mass, thermodynamic, and energy constraints[9]. Regulatory flux-balance analysis simply adds another set of constraints that control regulatory network. Integrated flux balance analysis (iFBA)[6] is another modification of FBA, which generates metabolic-regulatory networks. In contrast to rFBA, which simulates steady state behavior of the metabolic-regulatory network, iFBA uses differential equations to model a subset of the regulatory interactions. In both these methods, gene-transcription factor (TF) interactions are defined using Boolean logic allowing for only two possible states. Thus, genes and, consequently, reaction rates (fluxes) can be either completely on or off. It is a significant simplification in the relationship between the transcriptome and metabolome[5]. However, the greatest challenge in the successful application of the rFBA/iFBA methods is the manual curation of all the gene-TFs interactions. An extensive literature search has to be performed not only for a particular organism but also for a specific condition since transcriptional regulatory network (TRN) varies depending on external factors such as media composition. Consequently, Boolean logic based methods were applied to model metabolic-regulatory networks of only few well studied organisms (e.g. *Escherichia coli*[7], *Sacharomyces cerevisae*[10], or *Bacillus subtilis*[11]).

Published in 2010, **p**robabilistic **r**egulation **o**f **m**etabolism (PROM[5]) algorithm is the first fully automated method to build genome-scale metabolic-regulatory networks. This approach integrates currently widely available high-throughput gene expression data into constraint-based modeling (FBA). It also introduces probabilities as a representation of gene-transcription factor interactions. PROM has been successfully used to reconstruct genome-scale metabolic-regulatory networks for *Escherichia coli* and *Mycobacterium tuberculosis*[5]. In addition, it performed better than Boolean-based methods in mimicking cell behavior of transcriptionally perturbed *Escherichia coli* strains under various conditions[5]. Most importantly, the nature of this algorithm allows for it to be applied to any organism especially those less studied. The initial PROM performance results are very promising; however, this algorithm has only been applied to two organisms as well as to limited amount of data. Thus, further analysis is required to explore PROM's strengths and weaknesses.

1.3 Scope of this project

This study evaluates the feasibility and performance of the PROM algorithm as a tool to generate genome-scale metabolic-regulatory networks. First, we will discuss the theory behind this method (Chapter 2), followed by an analysis of the results presented in the original PROM paper (Section 3.1). In addition, a detailed analysis of the effect each of PROM's required components (metabolic model, transcriptional regulatory network and gene expression data) has on its performance will be presented. Chapter 3 will examine the impact of the metabolic model and experimentally determined transcriptional regulatory network on growth

predictions generated via PROM. In addition, new data sets will be utilized to assess accuracy in phenotype predictions for *E. coli* and *S. cerevisae*. In Chapter 4, we will include an analysis of the impact of the amount of gene expression data on the accuracy of PROM's gene essentiality and growth predictions. In addition, we consider the use of transcriptional regulatory networks inference algorithms, ASTRIX, and its effect on PROM simulations. Chapter 5 will present the application of PROM for flux predictions in *E. coli* and *S. cerevisae* mutant strains. We will conclude this thesis by summarizing and assessing PROM's performance for all considered applications in Chapter 6.

2. Probabilistic regulation of metabolism (PROM) algorithm theory

2.1 Factors required for the PROM algorithm

An overview of the PROM analysis process is depicted in Figure 1 below. The genome-scale metabolic-regulatory network can be build when the following data for a particular organism are available[5]: (1) reconstructed genome-scale metabolic network; (2) regulatory network structure (gene-transcription factor interactions); (3) gene expression data measured under various environmental and genetic perturbations. A detailed discussion of each data type will be included in the following Sections 2.1.1 – 2.1.3.

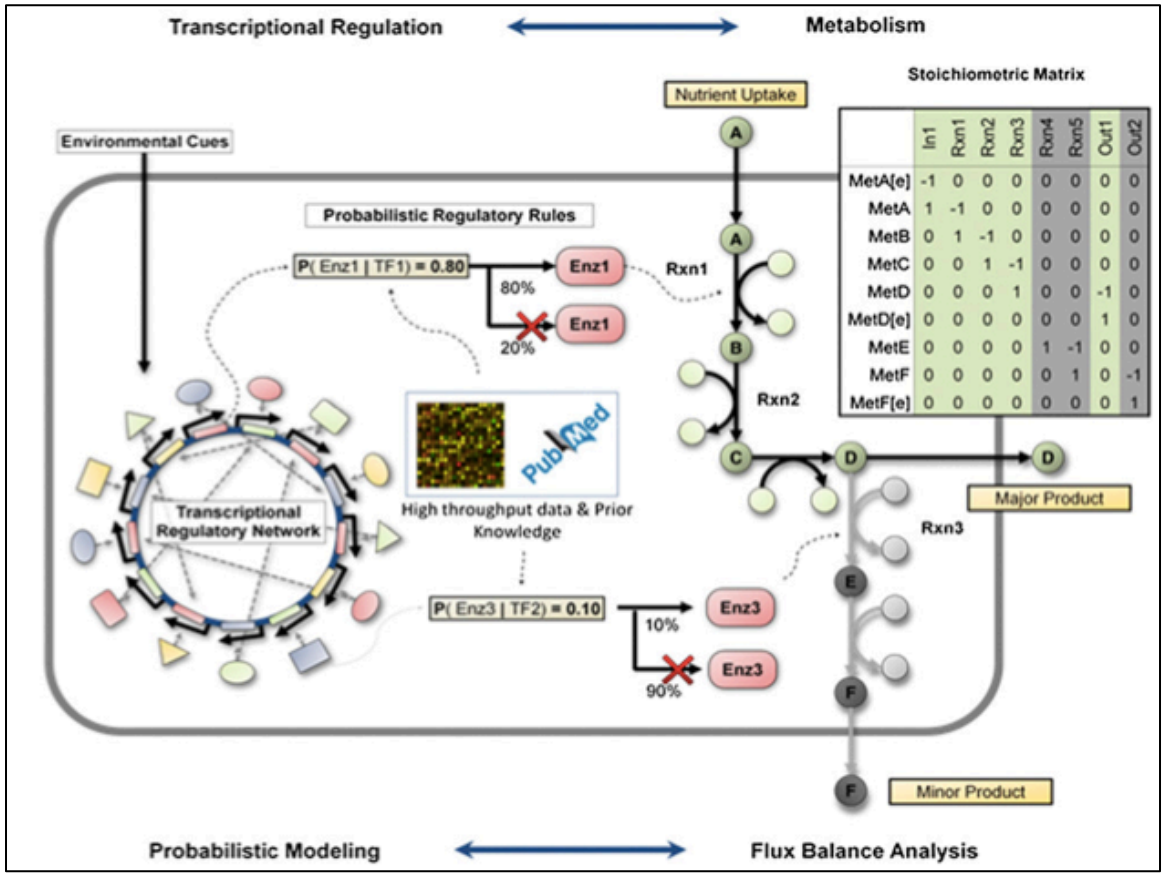

Figure 1 Overview of the process used to integrate the metabolic and regulatory network using PROM[5].

2.1.1 Genome-scale metabolic model

The number of published reconstructed genome-scale metabolic networks has been growing exponentially since 1999[3]. Currently, the reconstruction of metabolic networks is a well-established process[2], which requires a known genome sequence and sufficient amount of biochemical and physiological information. A well-defined four-step process has been used to manually generate all high-confidence metabolic models[3]. The procedure is as follows[3]: (1) create a draft reconstruction from genome sequence and gene annotation data (an automated process performed using various online databases such as KEGG); (2) manually curate the original reconstruction via primary literature search and translate it into a mathematical genome-scale metabolic model; (3) use detailed physiological data to validate the model; (4) continue improving accuracy of the model via additional experimental studies. As the last step indicates, published genome-scale metabolic models are subjects of continuous improvement. Thus, several versions of those reconstructions are available online for a specific organism. In this project it was assumed that, if needed, a metabolic model reconstruction could be generated for any organism. Consequently, all published reconstructions were used in the performed analysis.

2.1.2 Transcriptional regulatory network (TRN)

In contrast to metabolic model, the reconstruction of regulatory networks is more challenging since regulation in a cell occurs at multiple levels such as transcriptional control of mRNA abundance and various post-transcriptional modifications[4]. The mRNA abundance is the easiest to measure experimentally; therefore, reconstructed regulatory networks are, in fact,

transcriptional regulatory networks (TRN). Those types of networks are reconstructed based on high-throughput protein-DNA interaction data (measurement of mRNA abundance), very often in combination with inference algorithms, which determine the gene-transcription factor (TF) interactions based on transcriptiomic and genomic data. It is important to remember that those networks are static and describe all possible interactions. However, at any given point in time only a subset of those interactions is active. In this paper, we will both apply published transcriptional regulatory networks to the PROM algorithm for considered organisms as well as examine an inference algorithm, ASTRIX[12], of such networks from gene expression data.

2.1.3 Gene expression data

PROM also requires gene expression data to determine probabilities of gene-transcription factor interactions. What is actually measured to determine activity of a specific gene is the amount of mRNA produced under genetic and environmental perturbations. Initially such measurement was performed using low-throughput methods such as Northern Blotting and Quantitative PCR, which quantify only one gene at the time. Development of high-throughput methods allowed the whole transcriptome analysis in which thousands of genes are analyzed at the same time. Currently, a high-throughput data is widely available for various organisms and can be downloaded online from websites such as Gene Expression Omnibus (GEO). In case of lack of data, additional experiments could be easily performed. Thus, PROM can be readily applied to many studied organisms. In this paper, we will utilize only previously published gene expression (microarray) data sets to build genome-scale metabolic-regulatory networks.

2.2 PROM algorithm framework

2.2.1 Determination of probabilities

The innovation of the PROM algorithm lies in the introduction of probabilities as means to describe gene states and interactions between genes and their regulators (TFs). For instance, probability of a gene A being on when the regulating transcription factor B is off is given by P(A=1|B=0); and, similarly, if both gene and TF are on, the probability can be defined as P(A=1|B=1). PROM calculates these conditional probabilities using microarray data. For example, to determine probability of gene A to be expressed when transcription factor B is knocked out, PROM screens through the whole set of gene expression data and counts how many times gene A was observed to be on while TF B was off. If a data set is diverse and large enough, a robust estimate of the probability of the interactions can be determined[5]. To model the effect of a gene knockout (KO) on the genome scale, all probabilities for a particular TF and its targets have to be determined.

2.2.2 Gene -TF probabilities as a new constraint

Probabilities determined from gene expression data are then used to restrict fluxes through reactions controlled by the target genes[5]. First, a maximum reaction rate (Vmax) is determined for each reaction for only the metabolic model using flux variability analysis (FVA)[13]. Then, the regulatory interactions are incorporated by constraining fluxes. Specifically, calculated maximum reaction rates are multiplied by their respective probabilities to obtain new

maximum flux rates. For example, if gene expression data shows that gene A is expressed in 80% of the samples when gene A is absent, then on average, the maximum flux through reactions controlled by gene A will be (0.8)*Vmax. Consequently, the upper bound for the flux in PROM is calculated as follows: P(A=1|B=0)*Vmax. These regulatory constraints are 'soft' constraints, thus can be violated to maximize an objective function (biomass function)[5].

2.2.3 Objective function calculations

PROM, similarly to FBA, solves a linear optimization problem ($max\ w^T v$) for a specific objective function (optimal growth) subject to the following constraints: (1) $S.v=0$ and (2) $lb \leq v \leq ub$ where S is the stoichiometric matrix, v is a flux vector representing a particular flux configuration, $max\ w^T v$ is the linear objective function, and lb and ub are the vectors containing respectively minimum and maximum fluxes through all reactions[5]. In contrast to FBA, PROM finds flux distribution that satisfy not only 'hard' constraints (mass, thermodynamic, and energy constraints), but also 'soft' constraints for the regulatory interactions. Specifically, another linearization problem is $\min(\kappa.\alpha + \kappa.\beta)$, which is subject to constraints $lb' - \alpha \leq v \leq ub' + \beta$ where $\alpha, \beta \geq 0$, lb' and ub' are maximum and minimum fluxes respectively calculated using transcriptional regulatory constraints, and both α and β depict deviations from those values; the value of κ represents the penalty for such deviations. In the following analysis, κ was chosen to be 1 based on the analysis presented in the original PROM paper[5]. The GNU Linear Programming Kit solver was used to solve this optimization problem.

3. Growth predictions

3.1 Results presented in PROM paper

In the original PROM paper, its algorithm's performance was examined using gene deletion studies performed for well-studied *E. coli* and less known *M. tuberculosis* organisms. The summary of data used to generate those genome-scale metabolic-regulatory reconstructions is presented in Table 1 below.

Table 1 Model features and accuracy in predicting KO phenotypes in each organism[5] as published in the original PROM paper.

Feature	E. coli		M. tuberculosis
Metabolic model	iAF1260[14]		iNJ661[15]
Metabolic reactions	2,382		1,028
Total genes in the model	1,260		691
Regulatory data	RegulonDB[16] (version 4.0)		Balazsi et al.[17]
Regulatory interactions	1,773		218
Microarrays	907		437
PROM objective function	'Ec_biomass_iAF1260_core_59p81M'		'biomass_Mtb_9_60atp_test_NOF'
Validation data set	1,875 growth phenotypes[18]	14 growth phenotypes[18]	30 TF KO[19,20,21]
Accuracy, %	**85**	**-**	**95**
Sensitivity, %	**70**	**-**	**83**
Specificity, %	**91**	**-**	**100**
Correlation	**-**	**0.95**	**-**

As shown in the table, three validation studies were performed to assess PROM's performance. For *E. coli*, two phenotype data sets were used. First, PROM's ability to predict growth phenotypes (lethality) under 125 conditions of 15 TFs KO was compared to the rFBA method. PROM gave 85% accuracy, which was better than the 82.5% accuracy obtained via rFBA. Thus, PROM results not only in better accuracy but it is also easier to apply since it's an automated

method. Another study for *E. coli* focused on PROM's capability to predict growth rates. Consequently, the growth rates for 14 TF knockouts were simulated computationally and compared to experimental values. In this study, the growth rate was reported on minimal media and in both aerobic and anaerobic conditions. The correlation between experimental and predicted values was 0.95, which is an indication of PROM's high ability to predict exact growth values. However, since growth rates were reported for two conditions that result in a great difference in growth rate, this result was not truly reflecting PROM's accuracy. Consequently, the aforementioned application of PROM will be further investigated in this report. The last analysis performed was for essentiality of 30 transcription factors in *M. tuberculosis*. It is a less studied organism; thus, such investigation could verify the hypothesis that PROM could be applied to any species. Presented results were very promising for extending the use of this algorithm to model other genome-scale metabolic-regulatory networks of other organisms. The accuracy of essentiality prediction was reported to be 95%. In addition, for six transcription factors for which essentiality data was not available, PROM was used to estimate possible effects of their knockouts. All those TFs KOs showed no growth; thus, it was concluded that they were candidate essential genes. Later in this study, we will verify the accuracy of that conclusion.

3.2 Updated metabolic model and regulatory network

Since PROM's first publication in 2010, the metabolic models for both *E. coli* and *M. tuberculosis* have been updated. Similarly, new transcriptional regulatory networks for these

organisms have been published. Thus, to examine the effect of these two components on PROM algorithm's performance, the original studies have been performed with the use of revised models. We will first discuss results for *E. coli* followed by those from the *M. tuberculosis* study.

3.2.1 *E. coli* study

Since the first metabolic network of *Escherichia coli* K-12 MG1655 was assembled in 2000, updates have been performed periodically based on new experimental data. The reconstruction used in the original paper was from 2007. In this analysis, we will use the iJO1366 reconstruction that was published 4 years later. This metabolic model was "(1) updated in part using a new experimental screen of 1,075 gene knockout strains, illuminating cases where alternative pathways and isozymes are yet to be discovered, (2) compared with its predecessor and to experimental data sets to confirm that it continues to make accurate phenotypic predictions of growth on different substrates and for gene knockout strains, and (3) mapped to the genomes of all available sequenced E. coli strains, including pathogens, leading to the identification of hundreds of unannotated genes in these organisms".[22] The important modification to the previous model (iAF1260) is the addition of the wild type biomass objective function to the iJO1366 reconstruction while the iAF1260 model has only core biomass function. Also, the regulatory interaction data has been updated from version 4.0 to version 7.0 in RegulonDB. The major change observed in RegulonDB version 7.0 is the addition of Gensor

Units, which are genetic sensory response units that links transcriptional regulation with signaling networks. Since, PROM considers only metabolic and regulatory networks, it was expected that new transcriptional regulatory network should not have a great effect on its performance. The comparison of data used in both the original PROM *E. coli* reconstruction and this study is summarized in Table 2 below.

Table 2 Comparison of data used to generate genome-scale metabolic networks in original version (PROM 2010) and updated version (2012) for *E. coli*.

Feature	***E. coli* (PROM 2010)**	***E. coli* (PROM 2012)**
Metabolic model	iAF1260[14]	iJO1366[22]
Metabolic reactions	2,382	2583
Total genes in the model	1,260	1366
Regulatory data	RegulonDB[16] (version 4.0)	RegulondDB[23] (version 7.0)
Regulatory interactions	1,773	2204
Microarrays	907	907
PROM objective function	'Ec_biomass_iAF1260_core_59p81M'	'Ec_biomass_iJO1366_WT_53p95M'; 'Ec_biomass_iJO1366_core_53p95M'

First, the *E. coli* study for 14 TF KOs growth rates was repeated for the following two cases: (1) iJO1366 metabolic model with core biomass objective function and RegDB7.0 regulatory data; and (2) iJO1366 metabolic model with wild type biomass objective function and RegDB7.0 regulatory data. The comparison to experimental data and previous results is summarized in Table 3.

Table 3 Summary of growth rates for E. coli determined experimentally and computationally via PROM and various metabolic models and transcriptional regulatory networks.

Culture	Experimental growth	PROM (iAF1260 and RegDB4.0)	PROM (iJO1366 CORE and RegDB7.0)	PROM (iJO1366 WT and RegDB7.0)
WT + O2	0.7100	0.7382	0.8523	0.8539
WT - O2	0.4900	0.3850	0.4424	0.4436
Δ arcA + O2	0.6900	0.7651	0.8962	0.8979
Δ arcA - O2	0.3800	0.3224	0.3797	0.3807
Δ fnr + O2	0.6300	0.5635	0.6503	0.1891
Δ fnr - O2	0.4100	0.2181	0.2611	0.0760
Δ fnr/Δ arcA + O2	0.6500	0.6596	0.7247	0.2113
Δ fnr/Δ arcA + O2	0.3000	0.2040	0.2469	0.0719
Δ appY + O2	0.6400	0.7152	0.8274	0.8290
Δ appY - O2	0.4800	0.3287	0.3849	0.3860
Δ oxyR + O2	0.6400	0.7876	0.9059	0.9076
Δ oxyR - O2	0.4800	0.3287	0.3849	0.3860
Δ soxS + O2	0.7200	0.7680	0.8538	0.0317
Δ soxS - O2	0.4600	0.3768	0.4207	0.0156
correlation	**-**	**0.9495**	**0.9438**	**0.477**

The correlation obtained for the updated metabolic model and TRN with a core objective function was comparable to previous results: 0.95 and 0.94 respectively. However, for the wild type biomass objective function, the correlation was cut in half (0.477). The following plots (Figure 2 - Figure 4) illustrate the relation between PROM predicted growth rates vs. experimental growth rates.

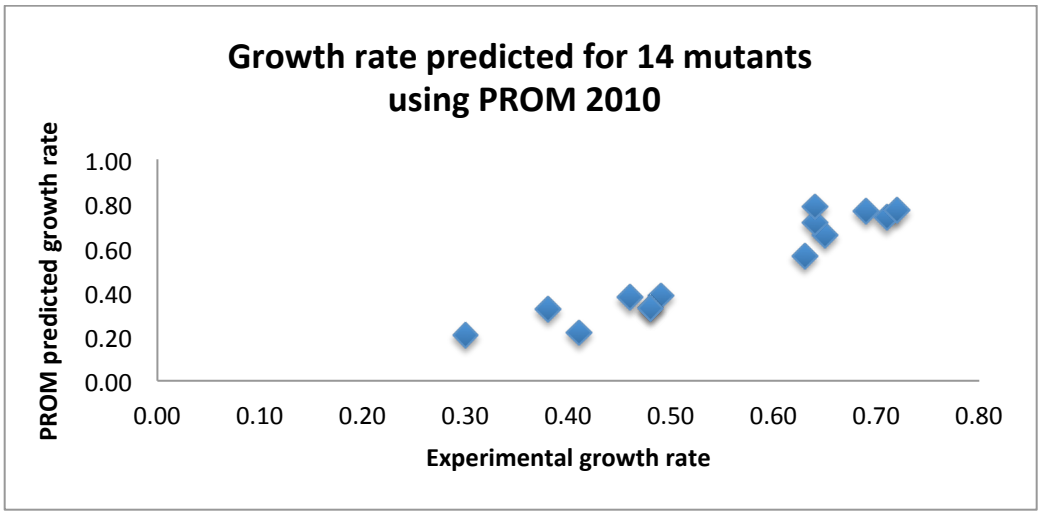

Figure 2 Comparison of the growth rate predicted by PROM vs. experimental growth rate for data used in orginal PROM.

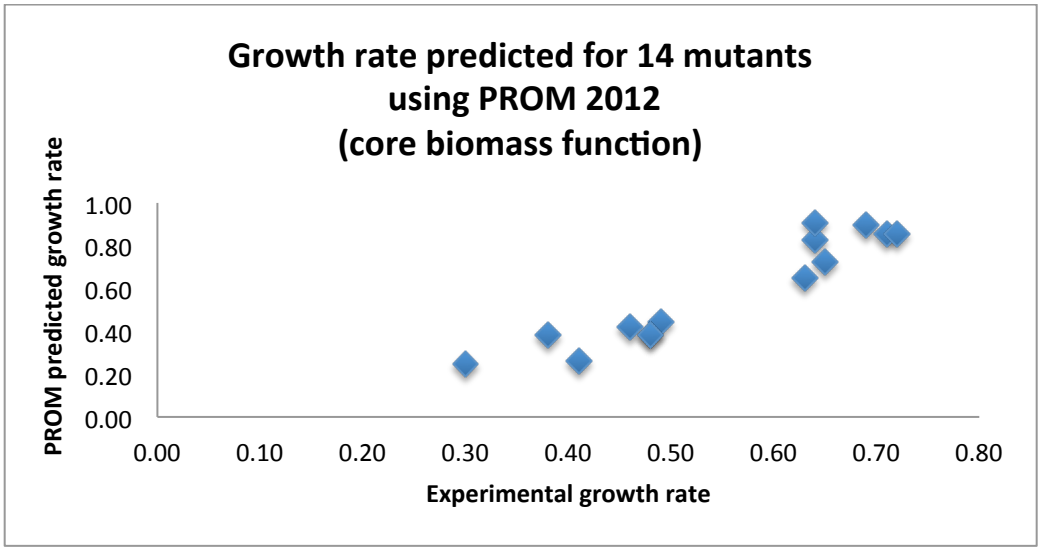

Figure 3 Comparison of the growth rate predicted by PROM vs. experimental growth rate for updated metabolic model and regualtory data. The objective function in PROM was set up to the core biomass function.

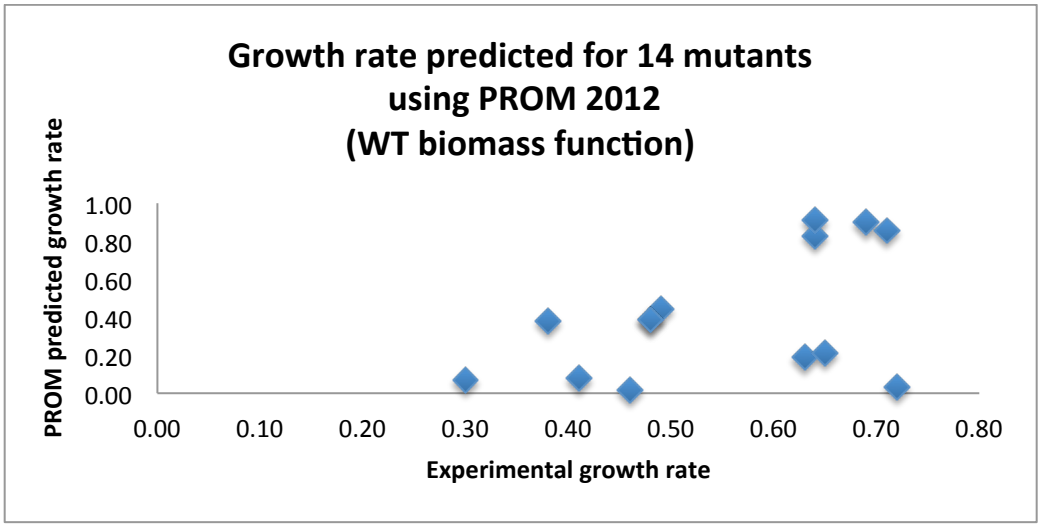

Figure 4 Comparison of the growth rate predicted by PROM vs. experimental growth rate for updated metabolic model and regualtory data. The objective function in PROM was set up to the wild type biomass function.

As seen from these plots and Table 3, PROM's results using the wild type biomass function incorrectly predict growth rates for the following gene knockouts: arcA; fnr and arcA; and soxS (each under both aerobic and anaerobic conditons). The double gene deletion (fnr and arcA) decreases the predicted growth rate. However, it is not the combination of both genes being

mutated, but only gene fnr is responsible for lowered growth rate. This was concluded from individual knockouts of fnr and arcA as presented earlier in Table 3. Thus, in this valiadation set only knockouts of genes fnr and soxS are incorrectly modeled using the wild type biomass function.

3.2.2 *M. tuberculosis* study

The previously used metabolic model (iNJ661) was updated to two new versions: iNJ661m (*in vitro* metabolic model) and iNJ661v (*in vivo* metabolic model). Both models were considered in this study. Also, an updated TR network reconstructed by Sanz et al.[25] was used as regulatory data. New more comprehensive TRN was created by adding novel characterized transcriptional regulations from 31 recent published experimental studies. A summary of the data used can be found in Table 4 below.

Table 4 Comparison of data used to generate genome-scale metabolic networks in original version (PROM 2010) and updated version (2012) for *M.tuberculosis*.

Feature	***M. tuberculosis* (PROM 2010)**	***M. tuberculosis* (PROM 2012)**
Metabolic model	iNJ661[15]	iNJ661m/iNJ661v[24]
Metabolic reactions	1028	1049
Total genes in the model	661	663
Regulatory data	Balazsi et al.[17]	Sanz et al.[25]
Regulatory interactions	218	714
Microarrays	437	437
PROM objective function	'biomass_Mtb_9_60atp_test_NOF'	'biomass_Mtb_9_60atp_test_NOF'

We repeated the gene essentiality study presented in the original PROM paper. The same 24 transcription factors were computationally knocked out and the metrics used to assess PROM's performance were accuracy, sensitivity and specificity. The following updated networks scenarios have been considered: (1) *in-vitro* metabolic model and Balazsi et al.[17] regulatory data; (2) *in-vivo* metabolic model and Balazsi et al.[17] regulatory data; (3) *in-vitro* metabolic

model and Sanz et al.[25] regulatory data; and (4) *in-vivo* metabolic model and Sanz et al.[25] regulatory data. The obtained examined metrics are presented in Table 5.

Table 5 Summary of PROM predictions for 24 TFs KO in *M. tuberculosis* using original and updated metabolic model and regulatory network.

Feature	PROM 2010	PROM (iNJ661m)	PROM (iNJ661v)	PROM (iNJ661m and Sanz TRN)	PROM (iNJ661v and Sanz TRN)
Accuracy %	96	96	88	83	75
Sensitivity %	83	83	83	83	83
Specificity %	100	100	89	83	72

The updated *in-vitro* metabolic model had no effect on PROM's performance. iNJ661 model was extended only by adding 21 metabolic reactions and 2 new genes. Thus, no change in accuracy of the genome-scale metabolic-regulatory model was expected. In contrast, *in-vivo* metabolic reconstruction (iNJ661v) decreased the accuracy of the growth phenotype predictions. This result was expected because gene expression data was collected *in-vitro* and all knockout studies were performed *in-vitro* conditions. The effect of the TRN reconstruction was examined next with each metabolic model. Using updated regulatory interactions data also decreased PROM's ability to predict growth phenotypes for 24 TFs KO. The new TRN was completely updated and included 3 times more gene-TF interactions when compared to the previous model. All interactions between regulators and their targets were determined experimentally; thus, we shall assume they were high-confidence interactions. In this case, we revised the gene essentiality assumptions made in the original PROM paper. Table 6 below indicates published data for all 30 TFs together with the original PROM prediction. Essentiality data was obtained from papers published by Sassetti et at. (2003)[19], Gao et al. (2005)[20], Lamichhane et al. (2003)[21], and Griffin et al. (2011)[26]. Table 6 presents curated literature essentiality data for 30 transcription factors and original PROM prediction. In the original paper,

6 transcription factors (Rv1359, argR, sigC, sigH, lrpA and Rv3575c) were analyzed without any essentiality data. PROM predicted those TFs to be essential. Thus, it was hypothesized that they were all candidate essential genes. In this study, we found that there exists experimental evidence that all those TRs are, in fact, nonessential. Consequently, the previously used set of 24 TF KOs phenotypes was incomplete and 30 TFs should be used instead.

Table 6 Gene essentiality experimental vs. PROM prediction data for 30 TFs.

Gene	PROM prediction	Literature
dnaA	essential	essential
Rv0485		
crp		
sigD		
kdpE		
ideR		
Rv1359	candidate essential	nonessential
argR		
sigC		
sigH		
lrpA		
Rv3575c		
oxyS	nonessential	nonessential
nadR		
hspR		
regX3		
Rv0586		
narL		
furA		
Rv1931c		
furB		
lexA		
pknK		
dosR		
birA		
sigF		
kstR		
cyp143		
embR		
sigE	nonessential	nonessential/essential

Incorporation of the new experimental evidence for essentiality of all 30 genes knockouts led to different results than those presented in the original paper. Table 7 summarizes these new results and shows that incorporation of new data lowers the originally reported accuracy and specificity results. In fact, accuracy drops from 96% to 77% and other metrics decreases from 100% to 75% for the original PROM set up. PROM's generated genome-scale metabolic model predicts 6 false positives, which indicates that it is missing some crucial biological functions. The study for the updated metabolic model (iNJ661m) gave better results than those from the iNJ661 metabolic model. We observed an increase from 77% to 80% and from 75% to 79% for accuracy and specificity respectively. Specifically, in contrast to the previous model, the new metabolic model correctly predicted gene Rv1359 to be nonessential. Therefore, it was observed that high-confidence metabolic model had an impact on PROM's performance.

Table 7 Summary of PROM results for revisited essentiality data for 30 TFs KO.

Feature	PROM 2010 (30 TFs)	PROM (iNJ661m) (30 TFs KO)	PROM (iNJ661m and Sanz TRN) (30 TFs KO)	PROM (iNJ661m and Sanz TRN) (58 TFs KO)
Accuracy %	77	80	73	70
Sensitivity %	83	83	86	75
Specificity %	75	79	70	69

Next, the same study was repeated for the new TRN network, which included a total of 58 transcription factors. The essentiality data for additional TFs was obtained from the same published sources as in the previous study. Table 8 shows experimental essentiality prediction together with PROM's prediction in which iNJ661m metabolic model and Sanz et al.[25] TRN were used. The overall accuracy for that genome-scale metabolic-regulatory reconstruction was of 73%, with sensitivity and specificity of 86% and 70% respectively. Since, a different validation

data set was used in this case, we cannot compare it to the previous results that considered only 30 TFs in contrast to 58 TFs. Again, this reconstruction predicts 12 false positive knockouts (the computational model indicates that the organism is not growing while experimental data shows the contrary). Thus, the new TRN misses some important biological relationships once again. Another explanation for such a difference between experimental and computational results could be the fact that various genes can be both essential and nonessential depending on different conditions. Therefore, PROM may model those knockouts assuming a different external environment even though it was specified to closely reflect experimental conditions. Also, there is no data available for the pyrR gene and the PROM model predicts it to be essential. Thus, we assumed that it could be a candidate essential gene.

Table 8 Gene essentiality experimental vs. PROM prediction data for 58 TFs.

Gene	Literature	PROM prediction
dnaA	essential	essential
Rv0485		
crp		
sigD		
ideR		
sigB		
phoP		
kdpE	essential	nonessential
mce3R		
Rv2017		
Rv1395	nonessential	nonessential
oxyS		
nadR		
hspR		
Rv0586		
narL		
furA		
Rv1931c		
furB		
lexA		
pknK		
dosR		
birA		
kstR		

Table 8 Gene essentiality experimental vs. PROM prediction data for 58 TFs.

Gene	Literature	PROM prediction
cyp143		
embR		
sigK		
Rv0465c		
Rv1359		
cmr		
higA		
Rv2021c		
clgR		
moxR3		
sigJ		
Rv3557c		
whiB4		
moxR2		
espR		
Rv0818		
argR	nonessential	essential
sigC		
sigH		
lrpA		
Rv3575c		
regX3		
sigF		
sigG		
mprA		
blaI		
Rv2034		
whiB3		
Rv3678c		
sigM		
sigE	essential/nonessential	essential
Rv0260c		nonessential
Rv0348		essential
pyrR	unknown	essential

The last analysis performed to reveal the impact of metabolic network reconstruction on PROM's performance was to use *in-vivo* iNJ661v model together with Balazsi et al.[17] TRN to model 30 TFs KOs. Predicted growth rates for all genes are presented in Table 9 for both *in-vivo* and *in-vitro* metabolic models. The only difference in growth rate was observed for the Rv1931c

gene. Further analysis showed that this gene is in fact *in-vivo* essential for optimal growth. Thus, we showed that PROM has potential for predicting *in-vivo* gene essentiality using microarray data for *in-vitro* conditions. Consequently, the metabolic model has a greater impact on PROM's performance than transcriptional regulatory network.

Table 9 Comparison of growth predictions obtained using PROM in combination with either *in-vitro* metabolic model (iNJ661m) or *in-vivo* metabolic model iNJ661v

Gene KO	PROM 2012 (iNJ661m)	PROM 2012 (iNJ661v)
'Rv0001'	0.030816647	0.022073171
'Rv0117'	0.052199236	0.040712737
'Rv0212c'	0.052199236	0.040712737
'Rv0353'	0.052199236	0.040712737
'Rv0485'	0.042063441	0.032807351
'Rv0491'	0.052199236	0.040712737
'Rv0586'	0.052199236	0.04066409
'Rv0844c'	0.052199236	0.040712737
'Rv1027c'	0.052199236	0.040712737
'Rv1221'	0.052199236	0.040712737
'Rv1267c'	0.052199244	0.040712737
'Rv1395'	0.052199244	0.040610479
'Rv1657'	0.035475191	0.039445729
'Rv1785c'	0.052199236	0.040712737
'Rv1909c'	0.052199236	0.040712737
'Rv1931c'	**0.052136843**	**0.039694779**
'Rv2069'	0.024049898	0.018757701
'Rv2359'	0.052199236	0.040712737
'Rv2711'	0.032240689	0.025146103
'Rv2720'	0.052199236	0.040712737
'Rv3080c'	0.052199236	0.040712737
'Rv3133c'	0.052199236	0.040712743
'Rv3223c'	0.047262311	0.040382257
'Rv3279c'	0.052199236	0.040712737
'Rv3286c'	0.052199236	0.040712737
'Rv3291c'	0.026099605	0.036132554
'Rv3414c'	0.036034294	0.028104922
'Rv3574'	0.052199236	0.040712737

3.3 New phenotype predictions

New growth phenotype studies for *E. coli* and *M. tuberculosis* have been published since PROM's original paper. Thus, new data became available for testing this algorithm's performance. Specifically, experimentally determined growth rates for *E. coli*[31] and *M. tuberculosis*[32] were available for 81 and 97 transcriptional regulatory mutants respectively. Detailed analysis of those data sets is presented in the remainder of this chapter.

3.3.1 *E. coli* study

Haverkorn et al.[31] analyzed growth rates for 91 transcriptional regulator mutants on glucose and galactose. Specifically, he reported growth rate (h^{-1}), hexose uptake rate (mmol $gCDW^{-1}$ h^{-1}) , and acetate secretion rate (mmol $gCDW^{-1}$ h^{-1}) for minimal media conditions. PROM's algorithm was used to simulate those studies and results are presented in Figures 5 through 9. The study was performed only for genes that were included in transcriptional regulatory network; thus, only 81 mutants were considered. The use of three different metabolic models for each growth condition was considered in order to investigate further impact of the metabolic network reconstruction on PROM's accuracy. The Balazsi at el.[17] TRN was used in this analysis for all considered genome-scale metabolic-regulatory reconstructions. The uptake rates for glucose or galactose, and secretion rate for acetate were set up in PROM to the exact values reported in the original experimental study.

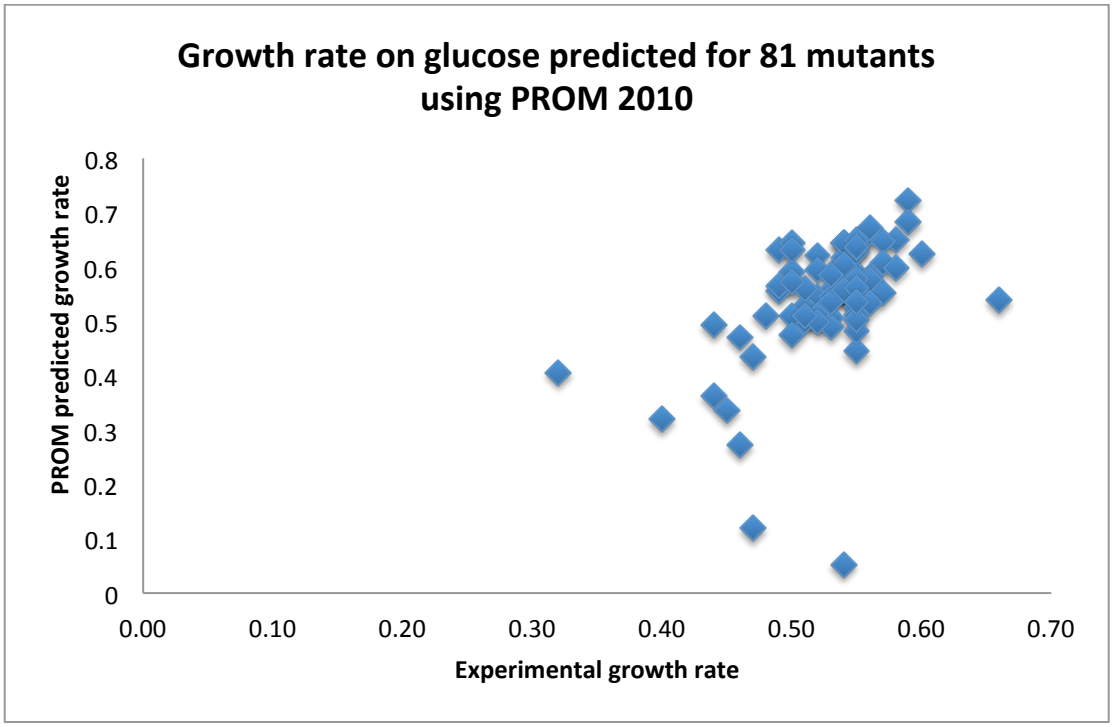

Figure 5 Comparison of experimental growth rate and computationally estimated (original PROM) growth rate for 81 TFs KOs of *E. coli* growing on glucose minimal media. Correlation of 0.50 was calculated using Matlab corr2 function.

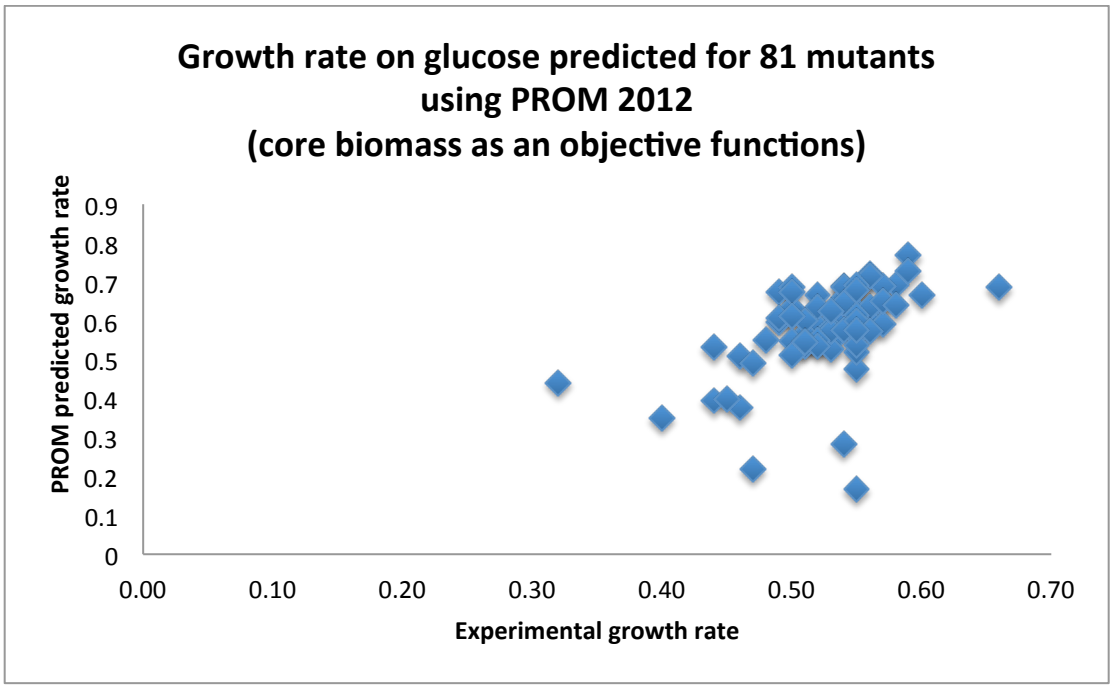

Figure 6 Comparison of experimental growth rate and computationally (PROM with updated metabolic model and core biomass objective function) estimated growth rate for 81 TFs KOs of *E. coli* growing on glucose minimal media. Correlation of 0.51 was calculated using Matlab corr2 function.

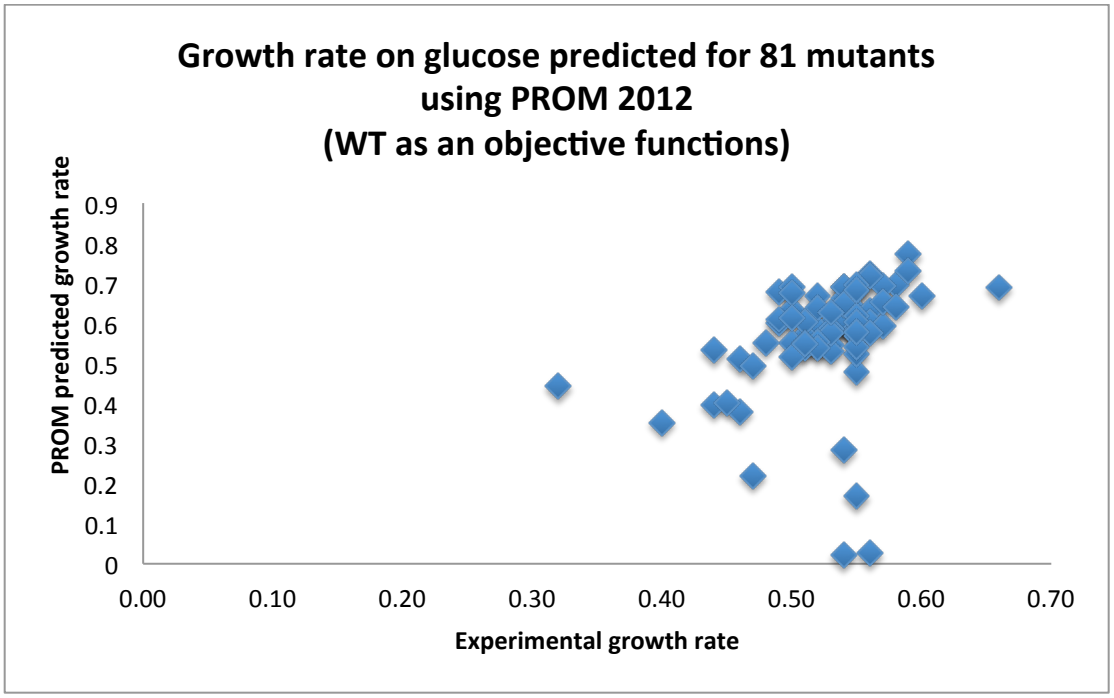

Figure 7 Comparison of experimental growth rate and computationally (PROM with updated metabolic model and wild type biomass objective function) estimated growth rate for 81 TFs KOs of *E. coli* growing on glucose minimal media. Correlation of 0.34 was calculated using Matlab corr2 function.

The correlations between experimental data and PROM predictions for growth on glucose were 0.50, 0.51 and 0.34 for each investigated case respectively. As shown in Figures 5 through 7 above, the data follows a straight-line trend with only few samples deviating considerably.

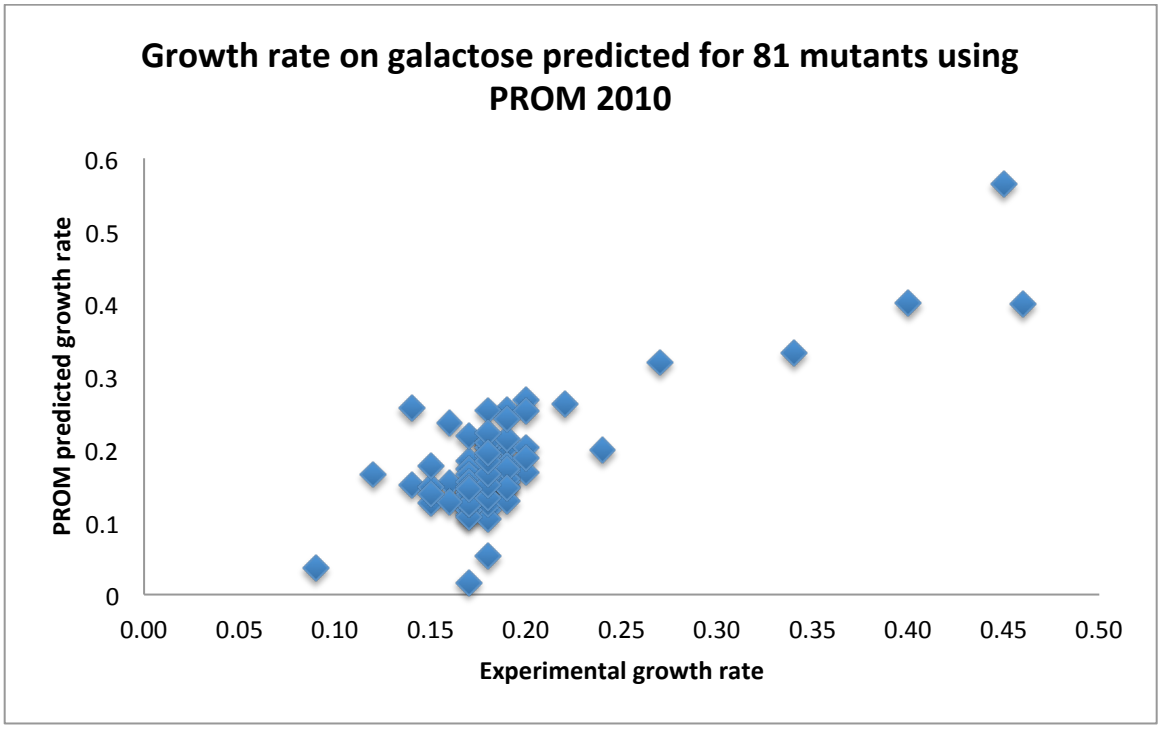

Figure 8 Comparison of experimental growth rate and computationally estimated (original PROM) growth rate for 81 TFs KOs of *E. coli* growing on galactose minimal media. Correlation of 0.81 was determined using Matlab corr2 function.

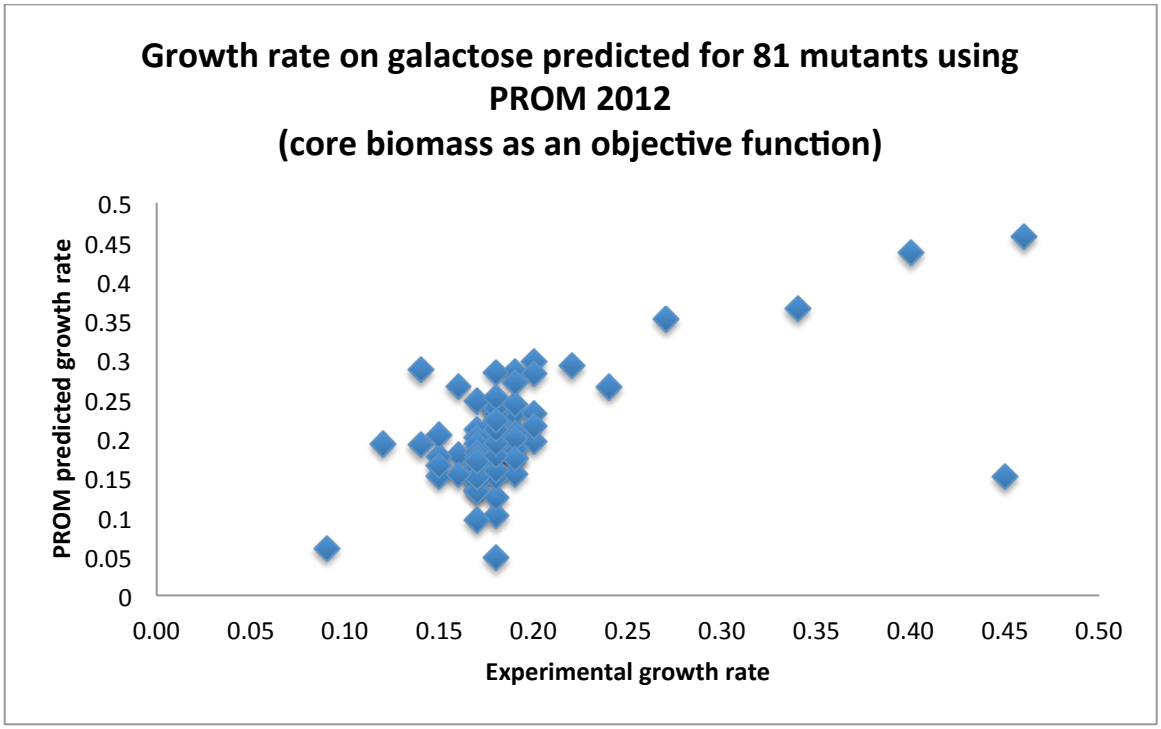

Figure 9 Comparison of experimental growth rate and computationally (PROM with updated metabolic model and core biomass objective function) estimated growth rate for 81 TFs KOs of *E. coli* growing on galactose minimal media. Correlation of 0.59 was determined using Matlab corr2 function.

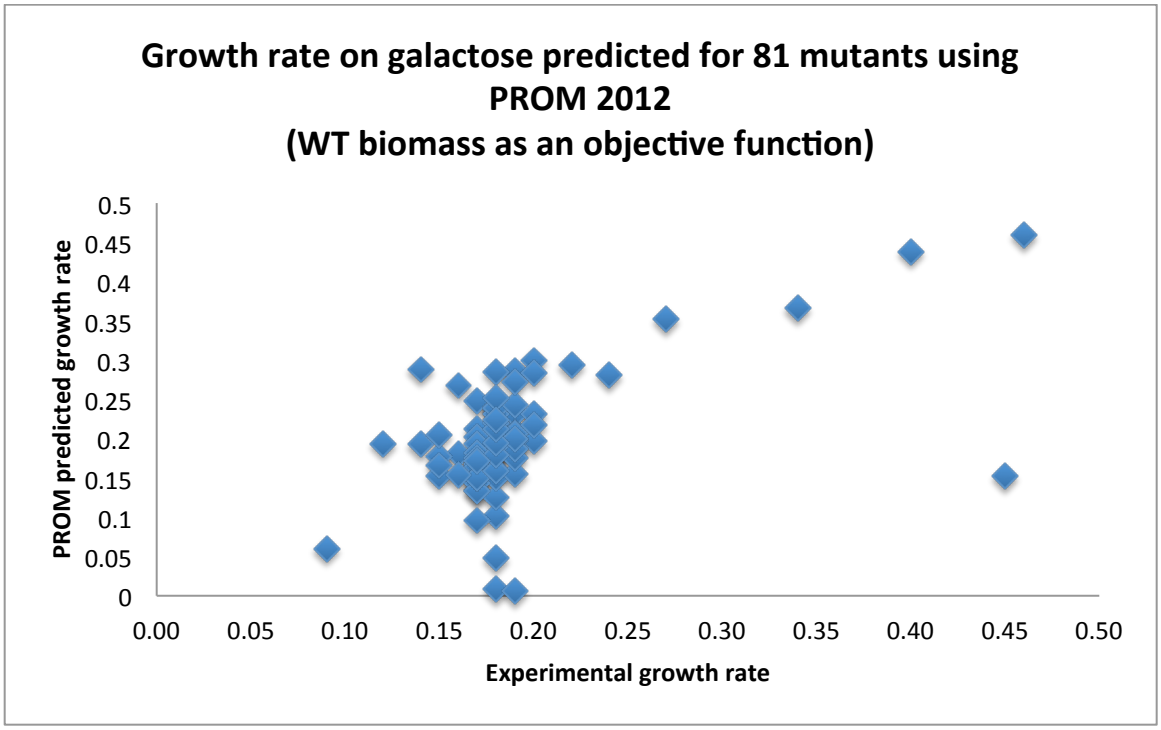

Figure 10 Comparison of experimental growth rate and computationally (PROM with updated metabolic model and wild type biomass objective function) estimated growth rate for 81 TFs KOs of *E. coli* growing on galactose minimal media. Correlation of 0.54 was determined using Matlab corr2 function.

A similar trend as for glucose was observed for galactose minimal growth conditions. The calculated correlations for experimental growth rate and PROM predictions were 0.81, 0.59 and 0.54 respectively for each case considered. Also, in this case, data is scattered along a straight line with few samples greatly deviating from the trend.

3.3.2 *S.cerevisae* study

Fendt et al.[32] analyzed growth rates for 119 transcriptional regulator mutants cultured on glucose, galactose and urea minimal media conditions. In this case, only growth rate (h^{-1}) was reported. The 97 TFs present in this study were part of the regulatory network. Thus, Figures 11

through 13 show results only for 97 mutants. To perform this analysis, PROM's algorithm was applied in order to reconstruct new genome-scale metabolic regulatory network for *S. cerevisae*. The following data was used for that process: (1) metabolic model reconstructed by Zomorrodi et al.[27]; (2) regulatory interaction data from YEASTRACT[28] online database; and (3) gene expression data from Many Microbe Microarrays Database. Also, uptake rates were adjusted to closely represent investigated minimal media conditions.

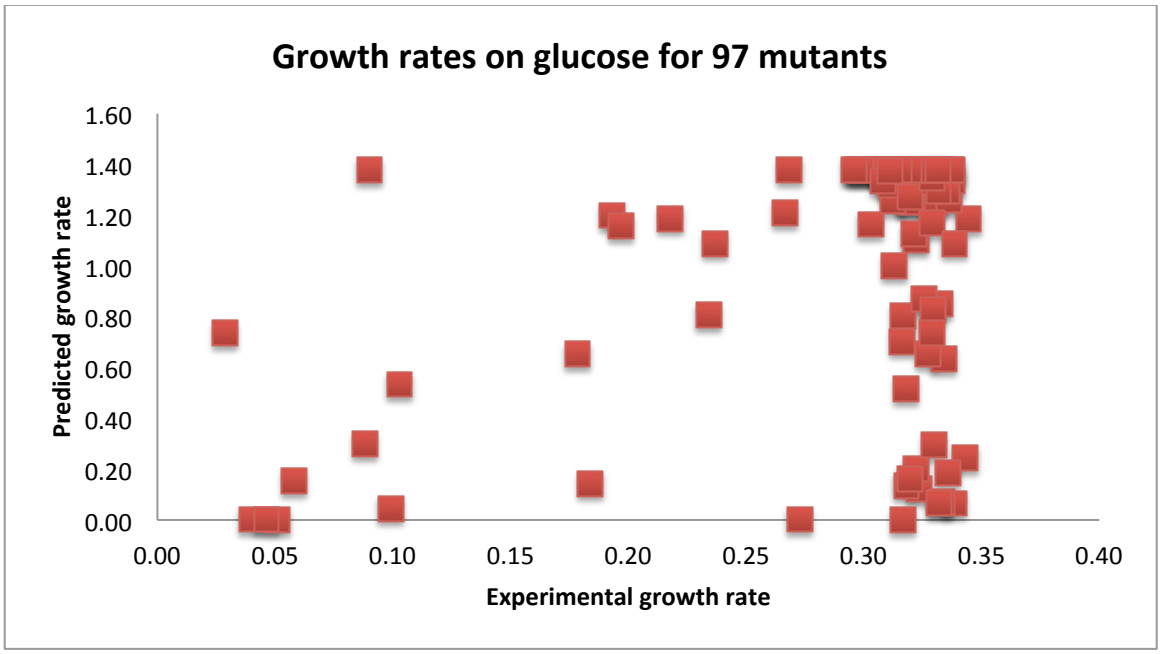

Figure 11 Comparison of experimental growth rate and computationally estimated growth rate for 97 TFs KOs of *S. cerevisae* growing on glucose minimal media. Correlation of 0.26 was determined using Matlab corr2 function.

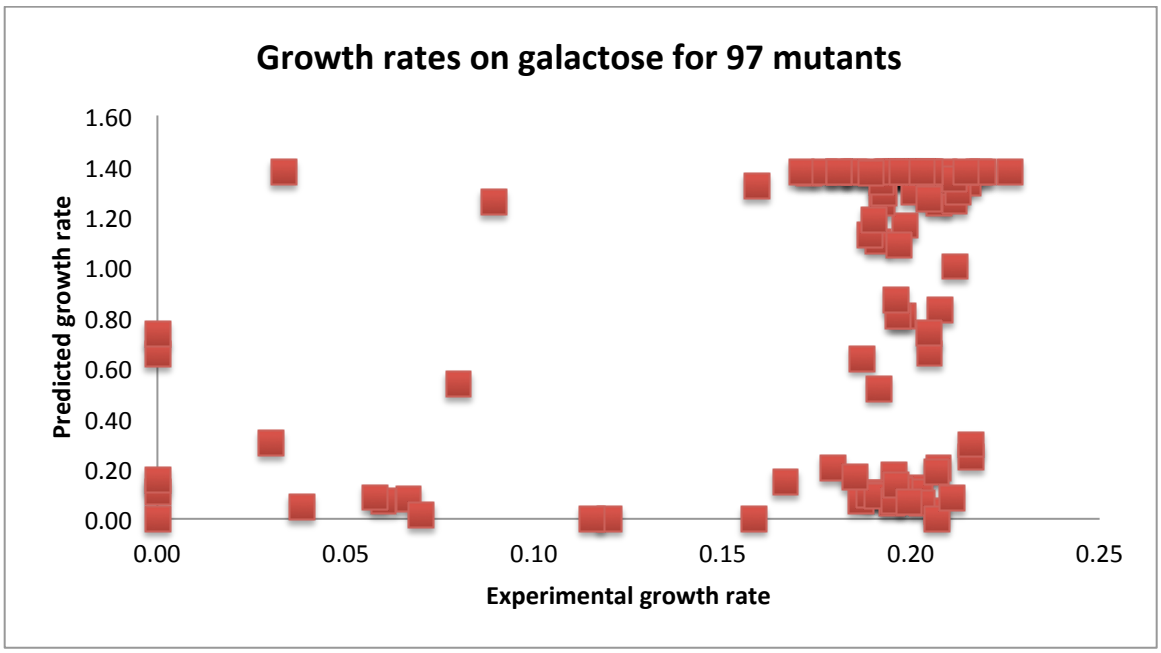

Figure 12 Comparison of experimental growth rate and computationally estimated growth rate for 97 TFs KOs of *S. cerevisae* growing on galactose minimal media. Correlation of 0.26 was determined using Matlab corr2 function.

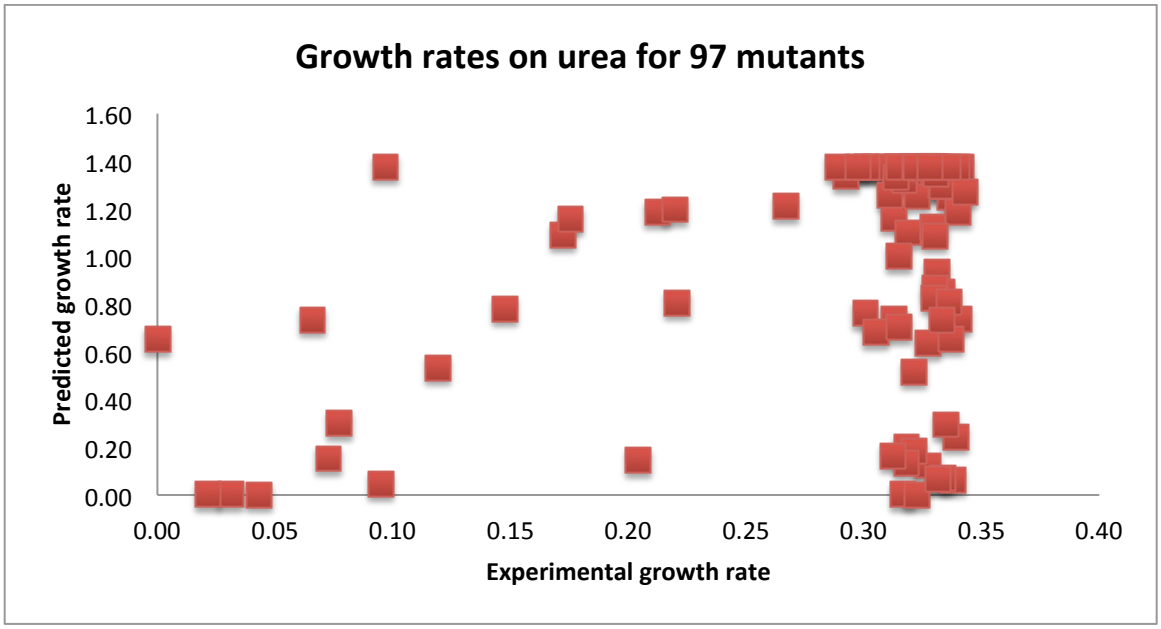

Figure 13 Comparison of experimental growth rate and computationally estimated growth rate for 97 TFs KOs of *S. cerevisae* growing on urea minimal media. Correlation of 0.24 was determined using Matlab corr2 function.

Correlations between experimental data and PROM's predicted growth rates for all considered cases are in the 0.24-0.26 range. Specifically, PROM determines that many of the

experimentally nonessential genes are lethal; thus, we can observe a vast majority of data points on the 0 growth rate line for predicted values. Application of PROM to *S. cerevisae* does not result in a successful high-confidence genome-scale metabolic-regulatory model. *S. cerevisae* (yeast), in contrast to *E. coli* and *M. tuberculosis*, is not a bacterium but an eukaryotic organism. Consequently, its complexity greatly exceeds other model organisms and PROM algorithm cannot decipher all the metabolic and regulatory interactions resulting in poor growth phenotype predictions.

3.4 Conclusions

In this chapter the effect of the accuracy of the metabolic model and transcriptional regulatory network reconstructions were examined. Specifically, updated models were applied to validation sets presented in the original PROM paper for *E. coli* and *M. tuberculosis*. For the most studied model organism - *E. coli*, we showed that further improvement of the metabolic model and addition of new regulatory interactions had no significant effect on PROM's performance. Thus, those networks used in the original reconstruction already had been high confidence. The analysis of two objective functions - core and wild type biomass - showed that PROM's algorithm successfully performs when maximizing the first one. The application of the later resulted in incorrect prediction that genes, fnr and soxS, were lethal. Similar analysis was performed for *M. tuberculosis*. In this case, the updated metabolic model also did not result in significant improvement in PROM's accuracy. However, the transcriptional regulatory network updated by Sanz et al.[25] lowered the predictivity of PROM. Since this TRN was generated based

solely on experimental data the validation set was explored further to investigate used essentiality data to verify that study had been performed correctly.

The first PROM paper included an analysis of 24 TF KOs for which essentiality data was available. We revised literature to determine essentiality for all 30 TFs present in TRN published by Balazsi et al.[17] This study showed that PROM's accuracy was 77% compared to 96% presented previously. In addition, it predicted 6 false positive gene knockouts. The incorporation of updated TRN with new validation data had the same impact as before: it lowered the model's accuracy.

Availability of *in-vivo* metabolic model for M. tuberculosis allowed us to explore the prospective use of PROM for prediction of in-vivo gene essentiality. Such capability would enable us to more accurately predict drug targets since occurring intracellular biochemical reactions depend strongly on the environment. In this case, the *in-vitro* gene expression data was available and combined with iNJ661v metabolic model and Balazsi et al.[17] TRN. Only one gene, Rv1931c, showed a different growth rate than while using iNJ661m (*in-vitro* model). The literature evidence confirmed that this gene was, in fact, *in-vivo* essential showing a prospective use of PROM in such analysis.

The exploration of new validation sets for prediction of TF KOs growth phenotypes using PROM showed that it can be applied only to less complex organisms. Analysis of *S. cerevisae* showed that PROM completely fails at simulating metabolic-regulatory interactions when applied to eukaryotic organisms resulting in experimental and predicted growth correlations in the range between 0.24-0.26. In case of E. coli, application of new data indicated that PROM is still

predictive; however, the correlations obtained for 81 transcriptional regulatory mutants were also lower than the validation sets presented in the original study. Instead of a correlation of 0.95, correlations of 0.5 were obtained. These new results reflect PROM's capability to predict growth rates better than the previous study since the initial data set included both aerobic and anaerobic conditions improving the correlation between experimental and computational data. The 81 mutant growth rates' study provides a better insight into PROM's capability to model genome-scale metabolic-regulatory networks.

4. Gene expression data and inferred TRN

4.1 Impact of amount of gene expression data on PROM

According to Chandrasekaran et al.[5], the PROM method can be easily extended to any organism for which a large number of gene expression data (microarray data) is available. However, in contrast to well-studied organisms such as *E. coli* and *M. tuberculosis* for which a great amount of microarray data is available, for a new organism gene expression experiments would have to be performed in order to use PROM. In this chapter, we will discuss if we can assess how many gene expression experiments should be performed for PROM's algorithm to be predictive.

PROM's performance was examined using gene deletion studies performed for two organisms: *M. tuberculosis* and *E. coli*. For each organism three experiments were performed: (1) transcriptional regulatory network was permuted; (2) the probabilities were randomly chosen in two ways (uniform distribution and permuted actual probabilities from the network); (3) number of input microarrays was varied for a network with defined probabilities and TRN. Depending on the availability of experimental data, PROM's predictions were examined by: (1) calculating accuracy, sensitivity, selectivity, and normalized score (when gene essentiality or lethality data was known) or (2) determining correlation between growth *in-silico* and *in-vivo* after each TF knockout.

4.1.1 *M. tuberculosis* study

The following statistics were used to measure PROM's performance for *M. tuberculosis*: accuracy, sensitivity, specificity, and normalized score. The obtained results are presented in Figures 14 through 17. The evaluation of PROM was based on model's ability to correctly

predict true positives (TP) and true negatives (TN) and to minimize false positives (FP) and false negatives (FN).

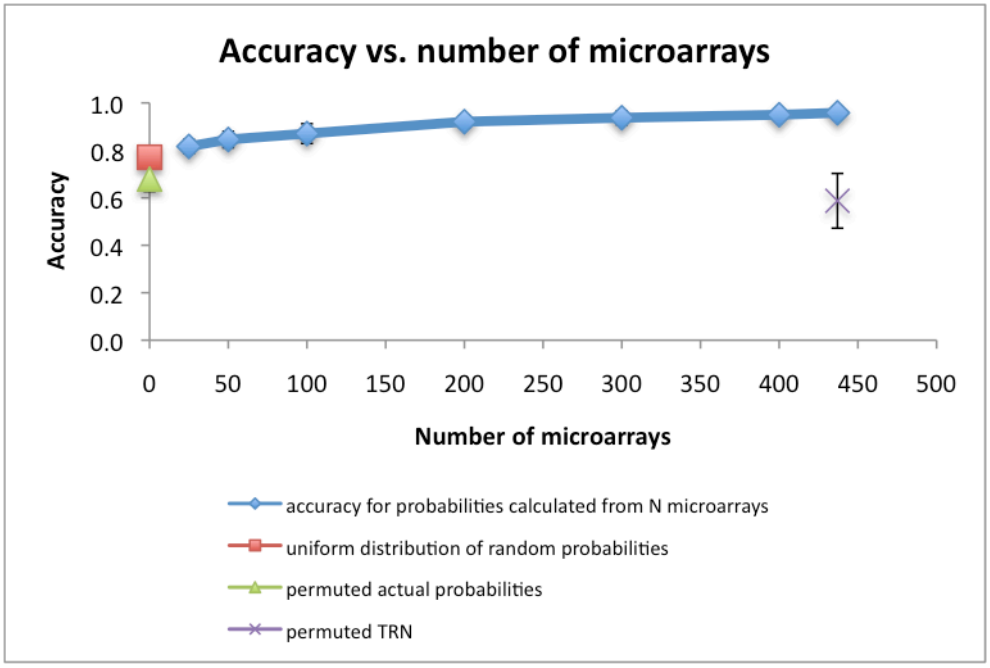

Figure 14 Effect of the number of microarrays on PROM's accuracy, where accuracy is defined as (TP+TN)/(TP+TN+FP+FN).

Accuracy was calculated for 4 cases as shown in Figure 14. First, the TRN was permuted and all microarrays were used resulting in an accuracy of ~0.59, which showed that transcriptional regulatory network has to be well defined for the model to be predictive. Next, instead of using microarray data to determine probabilities, those numbers were just assigned randomly either by using uniform random distribution or permuting actual probabilities from the model. In the first case the accuracy was 0.77 and in the later 0.68 showing that the type of connections in the network have an impact on PROM's performance. However, both results were lower than while using microarrays showing that gene expression data aids in PROM's accuracy. Last, the

number of microarray data was varied from 25 to 437, which increased the accuracy from 0.82 to 0.96. As seen in Figure 14, the effect of the number of microarrays has a great impact initially and then flattens out to a maximum value as expected.

Specificity and sensitivity were also calculated for the 4 cases described above. The obtained trends are presented in Figure 15 and Figure 16. The specificity (Figure 15) for permuted TRN as well as for random probabilities is significantly lower than when microarrays are used. In addition, incorporation of any gene expression data gives a high specificity (lowest measured 0.97). In contrast, the sensitivity analysis (Figure 16) yielded unexpected results showing that randomly chosen probabilities from a uniform distribution give the same sensitivity as the PROM model when all microarrays are being used. Such high sensitivity was a result of overmedication of false positives resulting in accurate measurement of true positives. In the case when the actual probabilities were permuted, the measured sensitivity was lower and equal to 0.72. Moreover, the permuted TRN also gave a significantly lower sensitivity than actual.

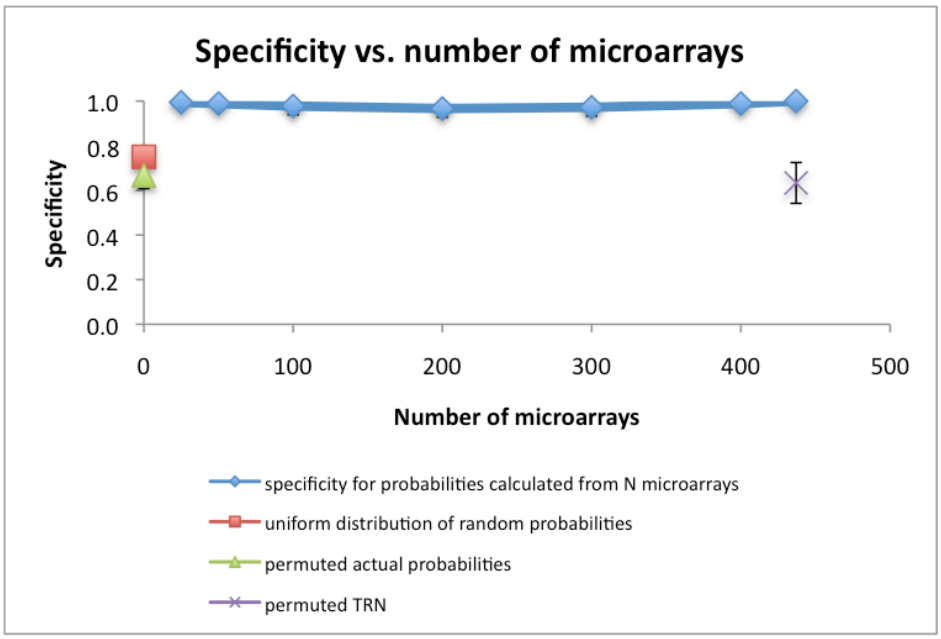

Figure 15 Effect of the number of microarrays on PROM's specificity, where specificity is defined as TN/(TN+FP).

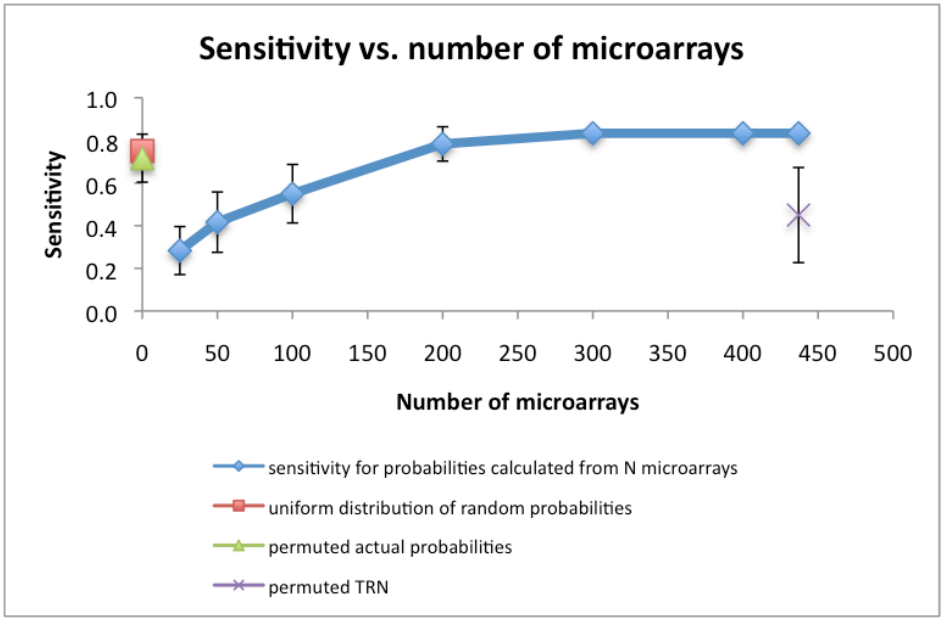

Figure 16 Effect of the number of microarrays on PROM's sensitivity, where sensitivity is defined as TP/(TP+FN).

Normalized score (Figure 17) was used to better measure PROM's performance for varying number of gene expression samples. That new statistics was used because the sensitivity analysis incorrectly indicated that zero microarrays performed equally well as fully characterized gene-transcription factor probabilities.

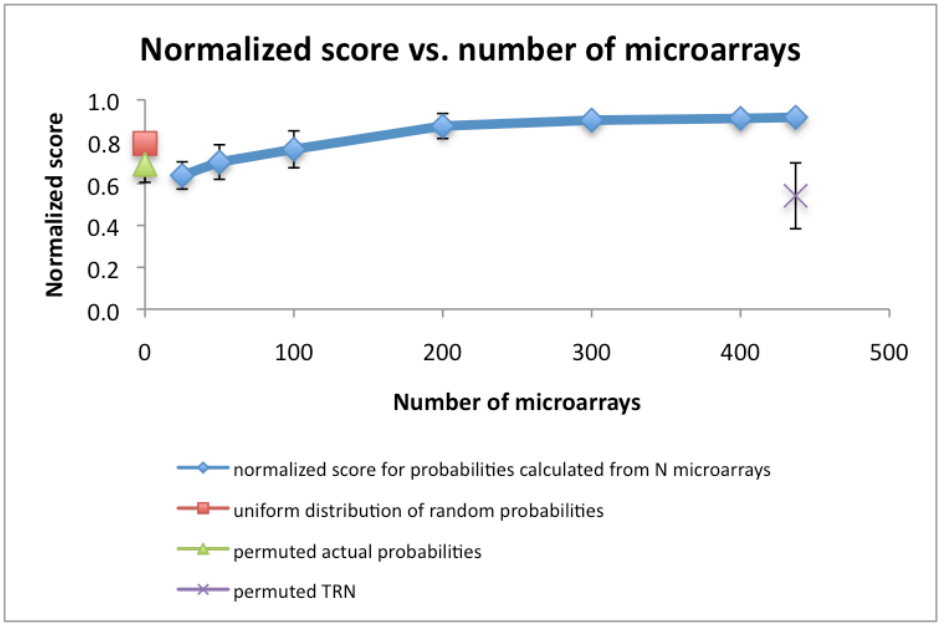

Figure 17 Effect of the number of microarrays on PROM's normalized score where it is defined as ½ specificity + ½ sensitivity.

4.1.2 *E. coli* study

PROM's performance for *E.coli* was assessed by measuring the correlation between growth rates predicted by the model and the experimental data. The obtained results for the 4 cases (as defined previously) are presented in Figure 18. The calculated correlation is significantly lower (~0.81) for randomized probabilities (zero microarrays) and slightly lower (~0.87) for the permuted TRN as compared to the maximum correlation of 0.95 obtained for the fully defined network. The interesting result of this study is a very small increase (0.04) in the correlation as

the number of microarrays is varied from 25 to 907. Initially, it was thought that this was a consequence of the degree of description of the network – *E. coli* as the best-defined organism. However, further investigation showed that the nature of the experimental data led to such a result. The given set of experimental growths was found for gene knockouts performed both at aerobic and anaerobic conditions. Consequently, two clusters of results were found and the correlation was plotted between them resulting in similar results thorough the whole range of gene expression data.

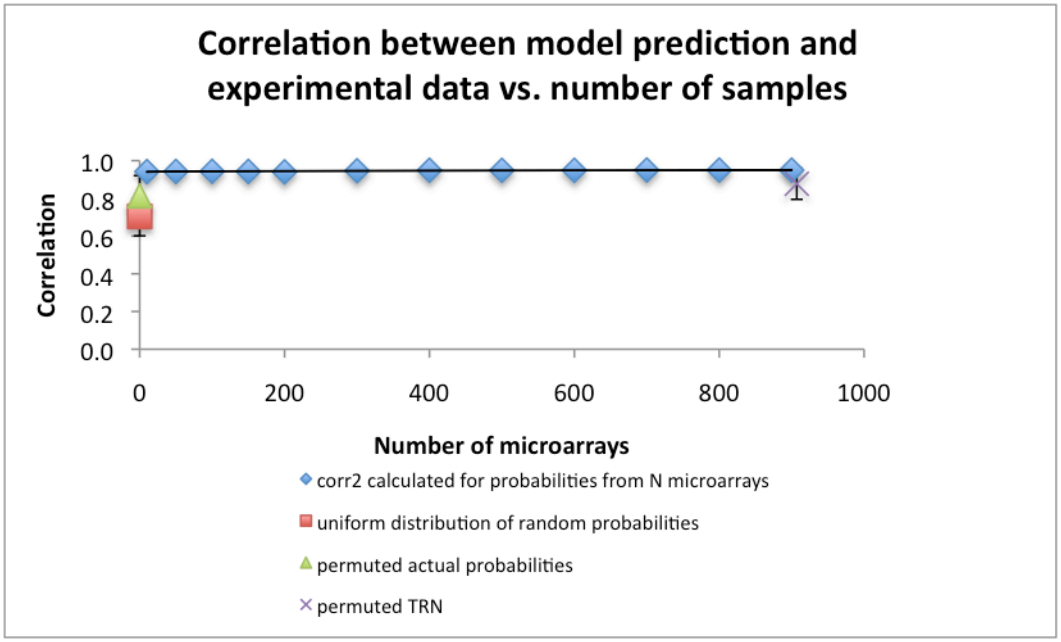

Figure 18 Correlation analysis for growth rate between PROM predictions and experimental data. Correlation was calculated using corr2 function in Matlab.

4.2 Use of inferred transcriptional regulatory networks and its effect on PROM

In the Section 4.1, we showed that TFs-gene interactions need to be well defined for PROM to be predictive. However, instead of using published TRN networks, in this section we will consider reconstructing the regulatory network using an inference algorithm. Once this is done, this reconstruction can be used as an input for PROM to model integrated genome-scale metabolic-regulatory network. This study will analyze how much experimental ChIP-chip data is required to establish TRN that can be successfully used by PROM.

The transcriptional regulatory networks for *M. tuberculosis* and *E. coli* were reconstructed using an approach called Analyzing Subsets of Transcriptional Regulators Influencing eXpression (ASTRIX)[12]. This algorithm is a combination of two well-known methods used for network inference: Accurate Reconstruction of Cellular Networks[29] (ARACNE) and Least Angle Regression[30] (LARS). ARACNE is first used to generate a network of high-confidence putative transcription factor-gene interactions which are further leveraged to predict expression in new conditions using LARS[12]. The data required by the ASTRIX algorithm to reconstruct TRN are gene expression data (microarrays) and a list of all known transcriptional factors.

4.2.1 *M. tuberculosis* study

ASTRIX was first applied to *M. tuberculosis* to reconstruct its TRN, which was later used as an input to PROM in order to simulate 30 TF KO growth phenotypes. The same gene expression

data as in previous studies was used while all known transcription factors were obtained from TubercuList as shown in Table 10.

Table 10 Summary of data used to reconstruct TRN using ASTRIX for *M. tuberculosis*.

Feature	TRN literature	TRN ASTRIX
Regulatory interactions	218	5340
Regulatory data	Balazsi et al.	TubercuList (transcription factors only)
Regulatory interactions present in both TRNs	2	
Microarrays	437	
Validation set	30 TF KOs	

The reconstructed TRN had 5340 gene-TFs interactions while the experimental TRN used in previous PROM studies had only 218 interactions (all present in metabolic model). However, there was only overlap of 2 interactions between both networks. Nonetheless, the ASTRIX TRN was still used to generate a genome-scale metabolic-regulatory model for *M. tuberculosis*. The accuracy for 30 TF KOs was only 0.54 when using inferred network in comparison to 0.96 when experimental network was used. The summary of accuracy studies performed for M. tuberculosis is presented in Figure 19 showing that inferred TRN gave the worst results out of all studies. Similar results were observed for the normalized score, which was 0.52 in comparison to 0.92 for fully experimentally determined regulatory interactions.

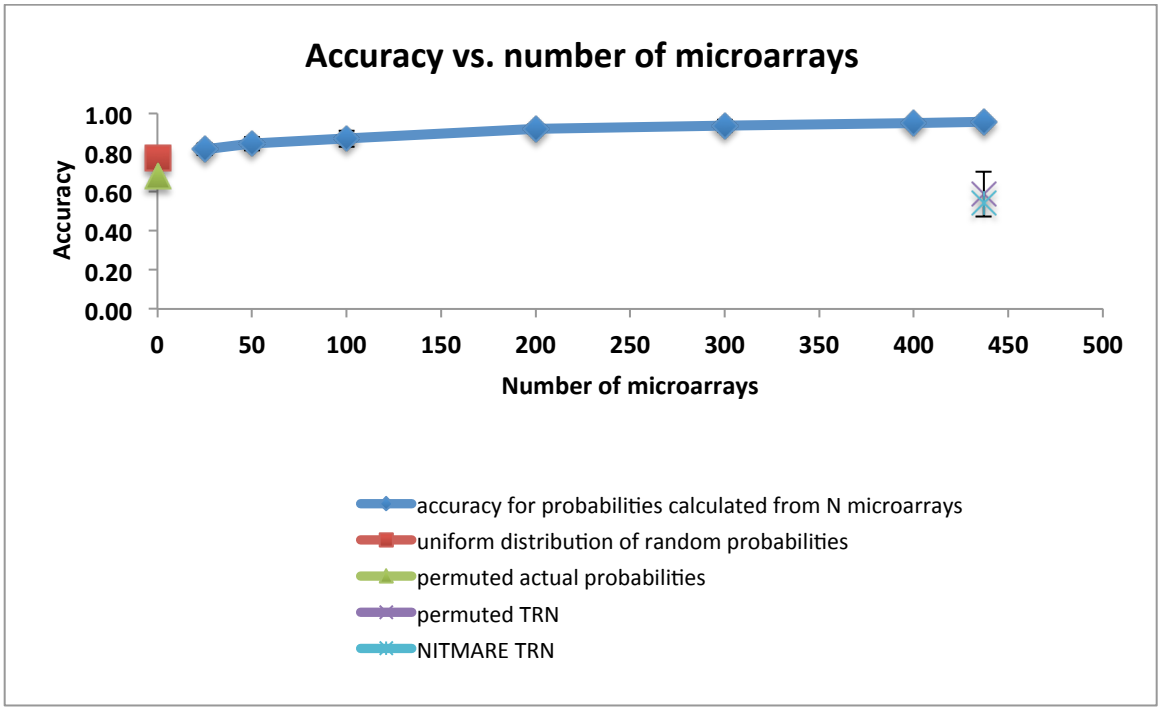

Figure 19 Summary of accuracy study results preformed for 30 TF KOs for *M.tuberculosis* for varying amount of gene expression data and confidence of TRN reconstruction accuracy.

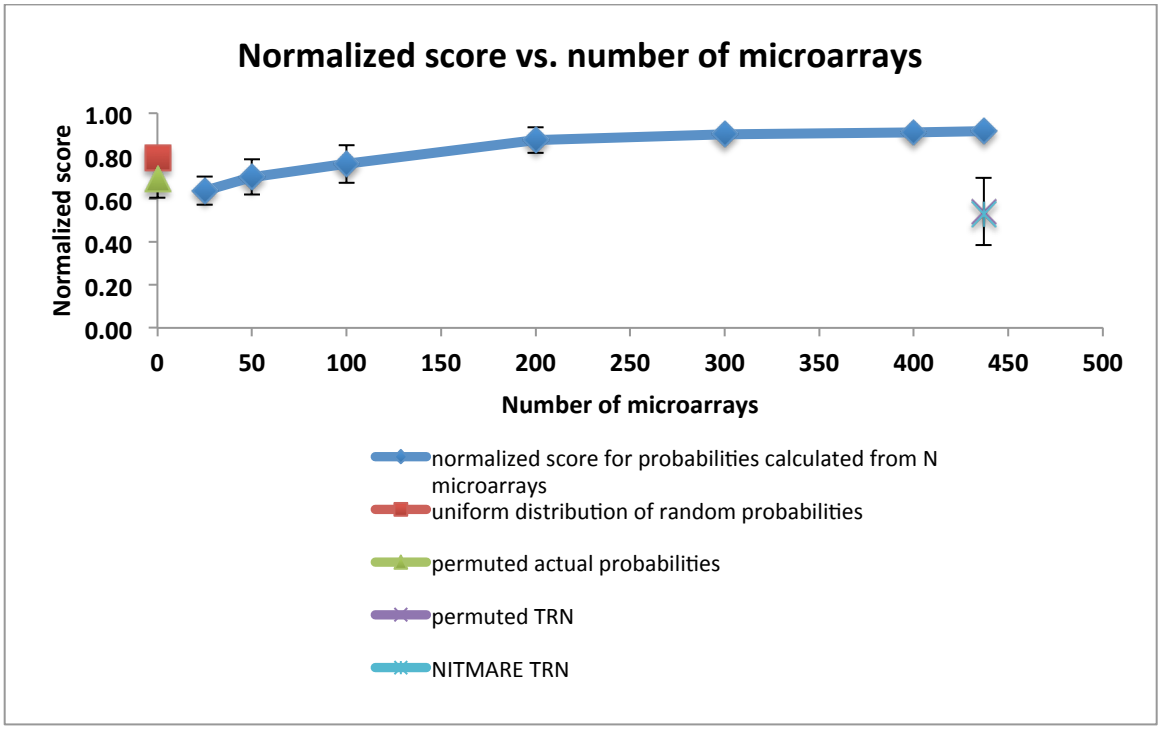

Figure 20 Summary of normalized score study results preformed for 30 TF KOs for *M.tuberculosis* for varying amount of gene expression data and confidence of TRN reconstruction accuracy.

4.2.2 *E. coli* study

An analogous study was conducted for *E. coli*. Table 11 summarizes the data used by ASTRIX. The same gene expression data (907 microarrays) was used as before. The list of all transcription factors was obtained from RegulonDB. In this case, 64 interactions that had been previously confirmed experimentally were also predicted by ASTRIX. However, there was only 3.06% overlap between both TRNs.

Table 11 Summary of data used to reconstruct TRN using ASTRIX for *E.coli*.

Feature	TRN literature	TRN ASTRIX
Regulatory interactions	1773	6841
Regulatory data	RegulonDB (version 4.0)	RegulonDB (version 4.0) (transcription factors only)
Regulatory interactions present in both TRNS	64	
Microarrays	907	
Validation set	14 growth phenotypes	

Next, inferred transcriptional regulatory network was used to generate a genome-scale metabolic model for *E. coli* via PROM's algorithm. In this case, the correlation between PROM's predicted growth phenotypes and experimental growth phenotype was 0.94 compared to 0.95 for fully experimentally defined TRN (see Figure 21). Such a small difference in correlation with only 3.06% overlap of inferred TRN with experimental data could indicate that the introduction of probabilities for regulatory interactions improves PROM performance significantly even when those interactions are not truly accurate. Additionally, the *E. coli* metabolic model reconstruction has high-confidence level since it has been updated several times since 2000 in contrast to the metabolic reconstruction for *M. tuberculosis*.

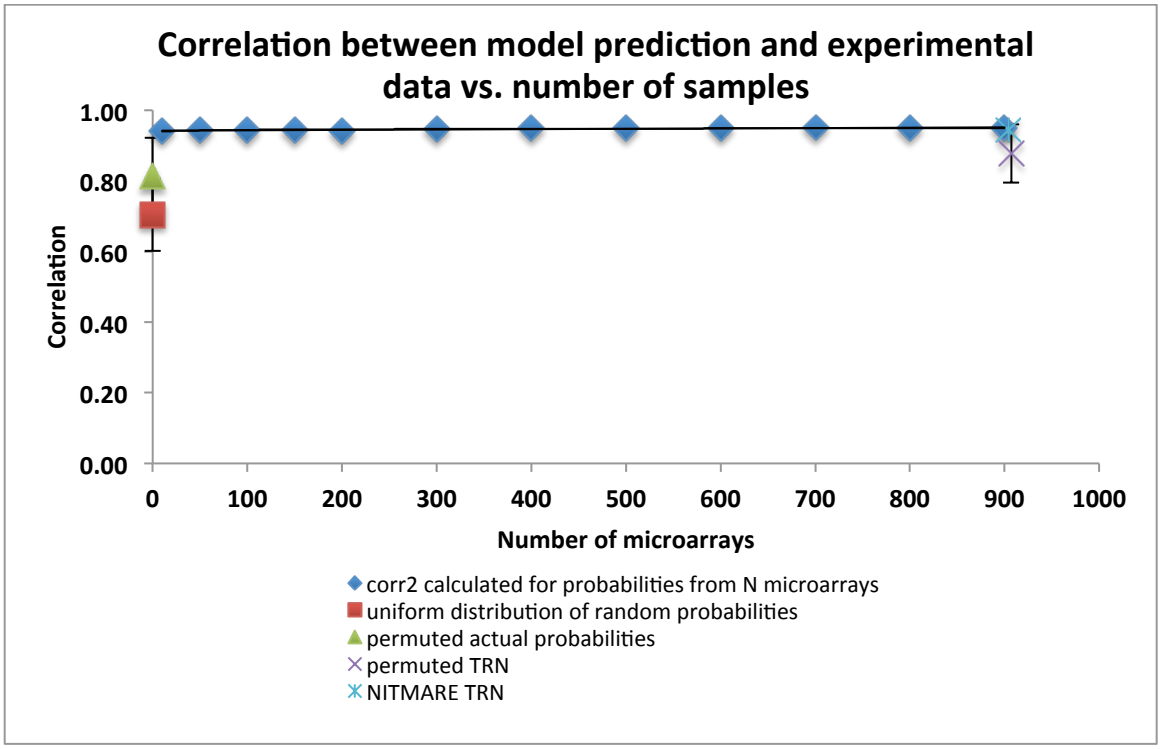

Figure 21 Summary of correlation study results preformed for 14 TF KOs growth phenotypes for *E.coli* for varying amount of gene expression data and confidence of TRN reconstruction accuracy.

4.3 Conclusions

The performed study demonstrated that PROM's performance depends on the degree of transcriptional regulatory network description as well as the amount of gene expression data. For the studied organisms, a well-defined TRN was required for PROM to be predictive. In addition, the incorporation of probabilities calculated from gene expression data improved PROM's performance. However, the impact of the number of microarray data used varied between those two organisms. For *M. tuberculosis* that effect was the greatest resulting in a 20% increase in accuracy as number of microarrays was varied from 25 to 437. In contrast, for *E. coli*, only a 4% increase in correlation was observed over the range from 25 to 907

microarrays. Such a difference could be a consequence of the metabolic and regulatory networks used for each model. *M. tuberculosis* is a less studied organism than *E. coli*; therefore, its networks are not as detailed as the ones of *E. coli*. Nonetheless, this study showed that the number of gene expression data required to reconstruct genome-scale metabolic-regulatory network using PROM can be assessed.

The TRN inference study resulted in the reconstruction of transcriptional regulatory networks for *M. tuberculosis* and *E. coli*. Both reconstructions had very small percentage of overlap with the experimentally determined gene-transcription factor interactions. Thus, ASTRIX should not be applied to reconstruct high-confidence TRNs. However, even though the obtained reconstructions were not accurate, the results gathered for each organism were very different. The *M. tuberculosis* study showed that poorly defined TRN results in low accuracy and normalized score for growth phenotype analysis. However, in the case of *E. coli*, transcriptional regulatory network had barely any effect on the correlation study. Thus, the accuracy of PROM's prediction depends on the balance between accuracy of the metabolic model and TRN.

5. Flux predictions *S. cerevisae* and *E. coli*

5.1 Importance of the distribution of intracellular reaction rates

Available genome-scale metabolic models and reconstructed transcriptional regulatory networks describe solely static interactions between transcription factors and their targets[31]. Those topological networks represent all (known up to date) possible relations between TFs and regulated genes. However, at any given time and conditions, only a subset of those interactions is active[31]. The expressed genes under particular condition can be determined by the measurement of the mRNA abundance. However, it is much harder to establish which of the occurring regulatory events control a specific metabolic function and to what extent. To answer that question, intracellular reaction rates (fluxes) have to be quantified and analyzed along with regulatory events to decipher how a specific metabolic process is controlled. So far, only two large-scale flux studies have been performed to reveal condition-specific networks of transcriptional regulation that control metabolic functions in *E. coli*[31] and *S. cerevisae*[32].

In depth understanding of control mechanisms would enable us to modulate cellular behavior. A cell operates via various biochemical pathways, many of which include alternative paths of obtaining the same phenotype. Thus, the exact knowledge of which reaction sequence takes place in a cell would yield answers to important questions in medicine and biotechnology. For example, it would be a breakthrough to induce apoptosis only in tumor cells without affecting healthy ones or to minimize formation of by-products in the process of producing a desired compound[32].

In this chapter we will explore the extension of PROM's algorithm to predict fluxes in addition to growth phenotypes. Flux Variability Analysis (FVA) was used to predict maximum and minimum reaction rates through all metabolic reactions after each gene knockouts. The direction of the flux change was examined and then compared to the experimentally reported flux ratio values for each study discussed below.

5.2 *S. cereviase*

Fend et al.[32] systematically quantified metabolic fluxes in 119 transcription factor deletion mutants of *Saccharomyces cerevisae* under five growth conditions[32]. Only 97 of these genes were included in the reconstructed genome-scale metabolic-regulatory model. Thus, only those TFs were considered in the further study. Also, PROM was applied to model only 3 out of the 5 considered growth conditions: glucose, galactose and urea minimal media while low pH and high osmolarity were omitted. The last two conditions cannot be simulated using PROM algorithm. Figure 22 represents the central metabolism and 6 flux ratios that were experimentally quantified using 13C-labeling method[33]. Reactions included in the metabolic model were searched through to determine reactions equivalent to presented flux ratios. Then, the overall change in the flux direction was determined (deviations less than 5% of initial flux ratio were assumed to be unmodified). Each flux ratio will be discussed individually in order to better present how they were simulated using PROM.

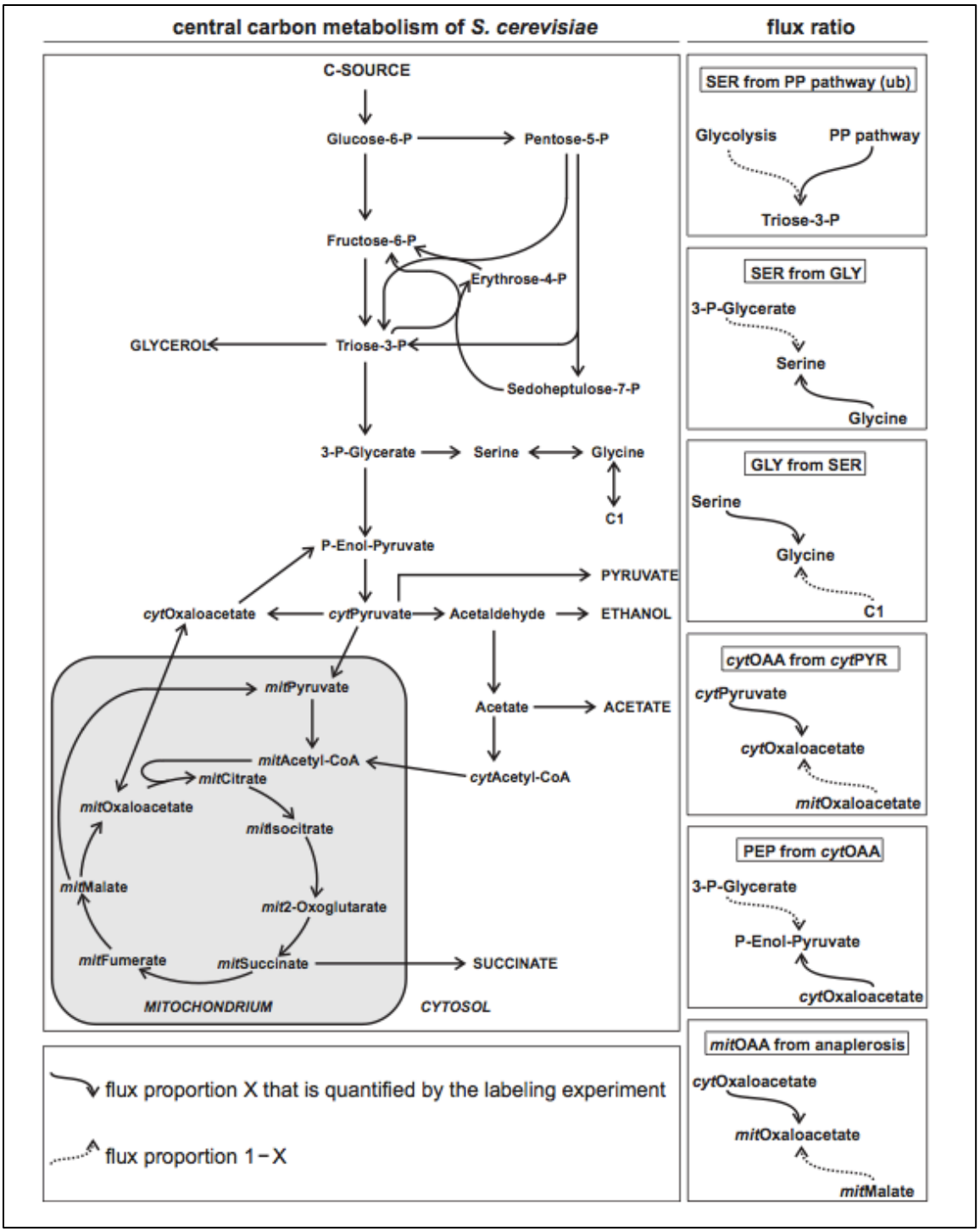

Figure 22 Representation of the central carbon metabolism of *S. cerevisae*. Flux ratios were calculated experimentally from 13C-labeled experiemnts as indicated above[32].

5.2.1 Serine from pentose phosphate pathway (SER from PP pathway)

Figure 23 below represents how 'SER from PP pathway' flux ratio (1:2) was determined experimentally. Thus, equivalent metabolic reactions present in the model were found based on information presented in KEGG database[34].

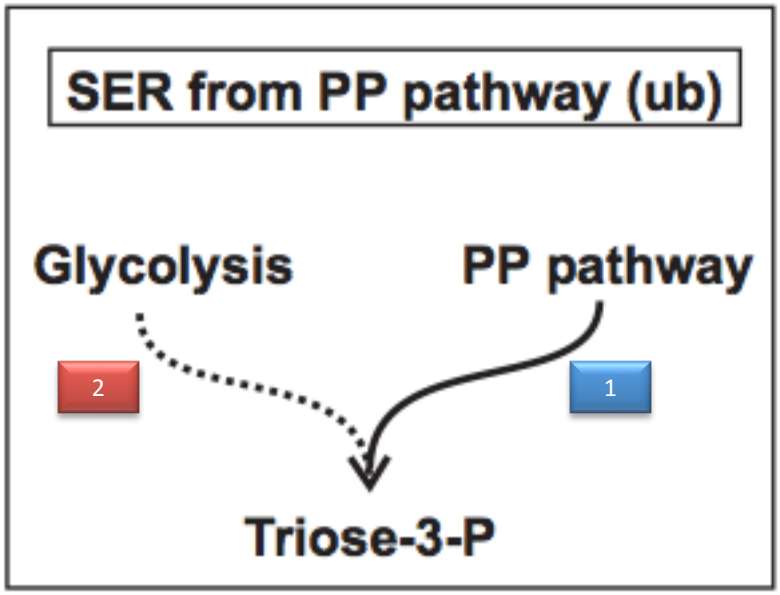

Figure 23 Illustration of how SER from PP pathway flux ratio was calculated.

Triose-3-P (known also as glyceraldehyde-3-P) from PP pathway reaction rate was calculated as follows: rxn1512 - rxn1483 + rxn1513; where reaction numbers refer to those present in *S. cerevisae* metabolic model. The triose-3-P generated through glycolysis was determined as follows: rxn741 + rxn1522; where reaction numbers also correspond to those in the metabolic model used. Table 12 presents flux formulas for each pathway and the full reaction formulas used for calculations. The flux ratio was obtained by dividing pathway 1 by pathway 2, and then the change in flux direction was determined.

Table 12 Summary of data used to calculate flux ratio 'SER from PP pathway'.

Pathway	Flux Formula	Reaction Formula
1: Triose-3-P from PP pathway	rxn1512-rxn1483+rxn1513	**rxns1512:** 'alpha_D_Ribose_5_phosphate + D_Xylulose_5_phosphate <=> Glyceraldehyde_3_phosphate + Sedoheptulose_7_phosphate ' **rxn1485:** 'Glyceraldehyde_3_phosphate + Sedoheptulose_7_phosphate <=> D_Erythrose_4_phosphate + D_Fructose_6_phosphate ' **rxn1513:** 'D_Erythrose_4_phosphate + D_Xylulose_5_phosphate <=> D_Fructose_6_phosphate + Glyceraldehyde_3_phosphate '
2: Triose-3-P from Glycolysis	rxn741+rxn1522	**rxn741:** 'D_Fructose_1_6_bisphosphate <=> Dihydroxyacetone_phosphate + Glyceraldehyde_3_phosphate ' **rxn1522:** 'Dihydroxyacetone_phosphate <=> Glyceraldehyde_3_phosphate '

PROM's algorithm correctly predicted 46 out of 97 flux changes on glucose, 33 out of 92 flux changes on galactose and 42 out of 96 flux changes on urea.

5.2.2 Serine from glycine (SER from GLY)

Analysis of serine from glycine flux ratio was determined in an analogous way to serine from PP pathway flux ratio. In this case, we found that only 2 direct reactions are required to transform 3-P-Glycerate to Serine and Glycine to Serine. Thus, only the flux ratio of those two reactions was required to compare with experimental results. Figure 24 presents how that flux ratio was calculated (pathway 1 divided by pathway 2) while Table 13 includes the details of this analysis.

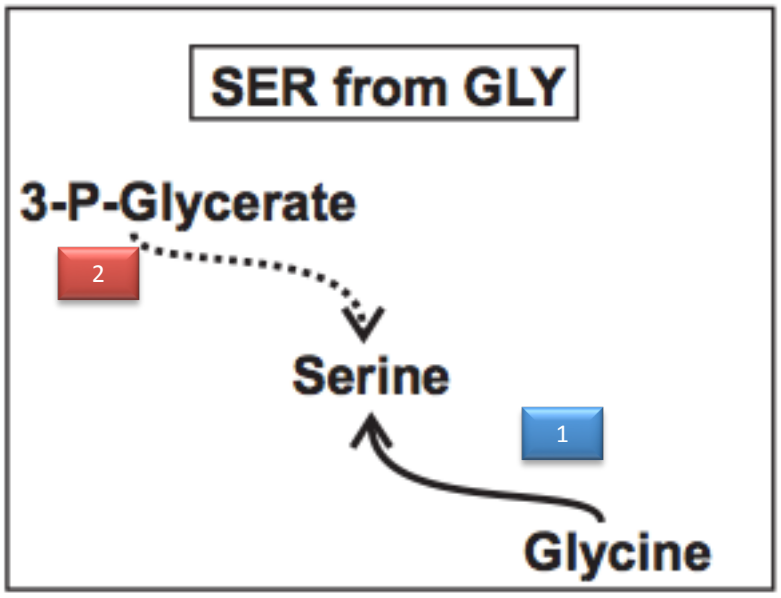

Figure 24 Illustration of how SER from GLY flux ratio was calculated.

Table 13 Summary of data used to calculate flux ratio 'SER from GLY'. Serine from 3-P-Glycerate was a series of three reactions (rxn1270, rxn1355, rxn1359), but all fluxes were equal; thus only one reaction was used for calculations.

Pathway	Flux Formula	Reaction Formula
1: Serine from Glycine	rxn827	'L_Serine + 5_6_7_8_Tetrahydrofolate <=> Glycine + H2O + 5_10_Methylenetetrahydrofolate '
2: Serine from 3-P-Glycerate	rxn1359	'H2O + O_Phospho_L_serine -> Phosphate + L_Serine '

PROM's algorithm correctly predicted 57 out of 97 flux changes on glucose, 45 out of 92 flux changes on galactose and 49 out of 96 flux changes on urea.

5.2.3 Glycine from serine (GLY from SER)

The analysis of flux ratio of formation of glycine from serine to glycine from carbon source is presented in Figure 25 and Table 14. In this case, the first reaction is determined as before (rxn827), and the other one is a series of three reactions, all of which have equal flux rates in PROM simulation. For this reason, only one reaction (rxn820) was applied to calculate the ratio.

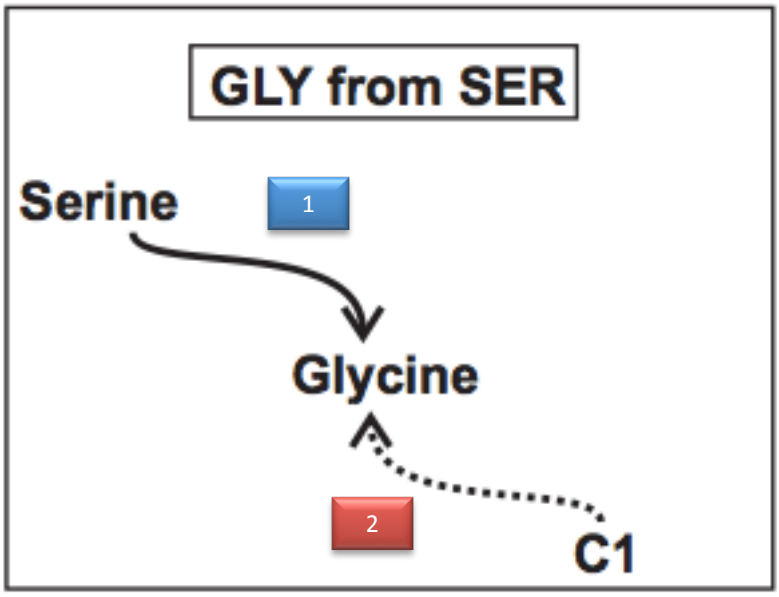

Figure 25 Illustration of how GLY from SER flux ratio was calculated.

Table 14 Summary of data used to calculate flux ratio 'GLY from SER'. Glycine from C1 was a series of three reactions (rxn818, rxn819, rxn820), but all fluxes were equal; thus only one reaction was used for calculations.

Pathway	Flux Formula	Reaction Formula
1: Serine from Glycine	rxn827	'L_Serine + 5_6_7_8_Tetrahydrofolate <=> Glycine + H2O + 5_10_Methylenetetrahydrofolate '
2: Glycine from C1	rxn820	'Dihydrolipolprotein + Nicotinamide_adenine_dinucleotide <=> H + Lipoylprotein + Nicotinamide_adenine_dinucleotide___reduced '

PROM's algorithm correctly predicted 65 out of 97 flux changes on glucose, 71 out of 92 flux changes on galactose and 49 out of 96 flux changes on urea.

5.2.4 cytOxaloacetate from cytPyruvate

Similar analysis was performed for cytOAA from cytPYR flux ratio. Only two direct reactions were required for this analysis. Figure 26 and Table 15 outline the performed analysis.

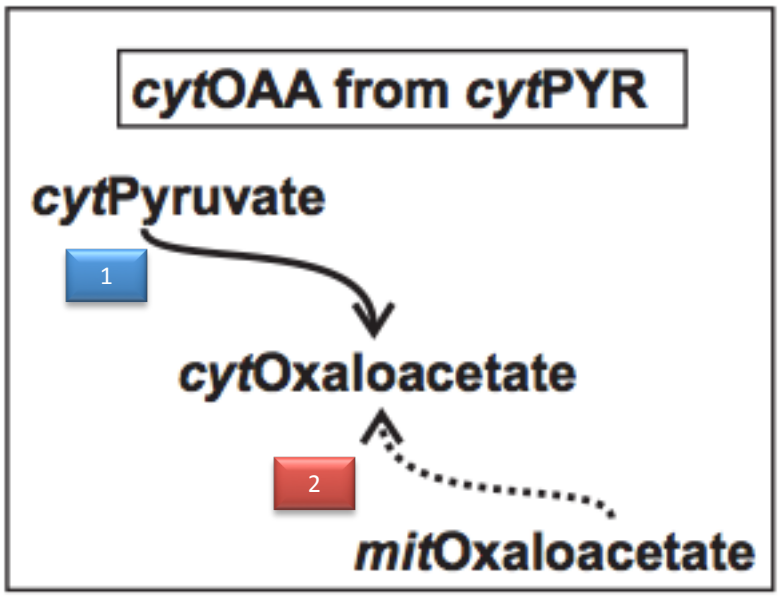

Figure 26 Illustration of how cytOAA from cytPYR flux ratio was calculated.

Table 15 Summary of data used to calculate flux ratio 'cytOAA from cytPYR'.

Pathway	Flux Formula	Reaction Formula
1: cytOxaloacetate from cytPyruvate	rxn1247	'ATP + Bicarbonate + Pyruvate -> ADP + H + Oxaloacetate + Phosphate '
2: cytOxaloacetate from mitOxaloacetate	rxn1211	'H + Oxaloacetate <=> H + Oxaloacetate '

PROM's algorithm correctly predicted 37 out of 97 flux changes on glucose, 27 out of 92 flux changes on galactose and 46 out of 96 flux changes on urea.

5.2.5 P-Enol-Pyruvate (PEP) from cytOxaloacetate

PEP is formed from 3-P-Glycerate through two reactions. Thus, the limiting reaction was chosen to calculate the flux ratio. In contrast, PEP from cytOAA is a one step reaction. Figure 27 and Table 16 present the outline of the analysis.

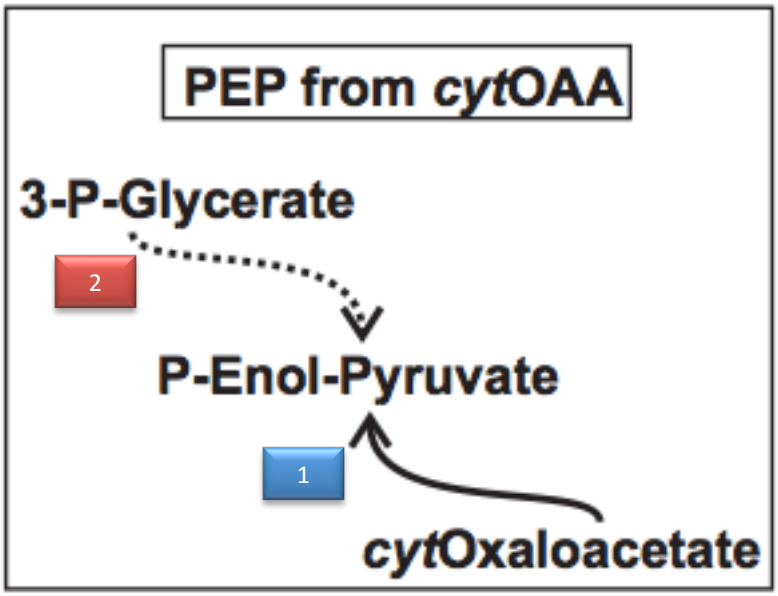

Figure 27 Illustration of how PEP from cytOAA flux ratio was calculated.

Table 16 Summary of data used to calculate flux ratio 'PEP from cytOAA'.

Pathway	Flux Formula	Reaction Formula
1: PEP from cytOxaloacetate	rxn1324	'ATP + Oxaloacetate -> ADP + CO2 + Phosphoenolpyruvate '
2: PEP from 3-P-Glycerate	min(rxn1274,rxn467)	**rxn1274:** 'D_Glycerate_2_phosphate <=> 3_Phospho_D_glycerate ' **rxn467:** 'D_Glycerate_2_phosphate <=> H2O + Phosphoenolpyruvate '

PROM algorithm correctly predicted 37 out of 97 flux changes on glucose, 27 out of 92 flux changes on galactose and 46 out of 96 flux changes on urea.

5.2.6 mitOxaloacetate from anaplerosis

In this scenario (Figure 28), two direct reactions (Table 17) were used to calculate flux direction change using PROM. PROM algorithm correctly predicted 63 out of 97 flux changes on glucose, 33 out of 56 flux changes on galactose and 46 out of 96 flux changes on urea.

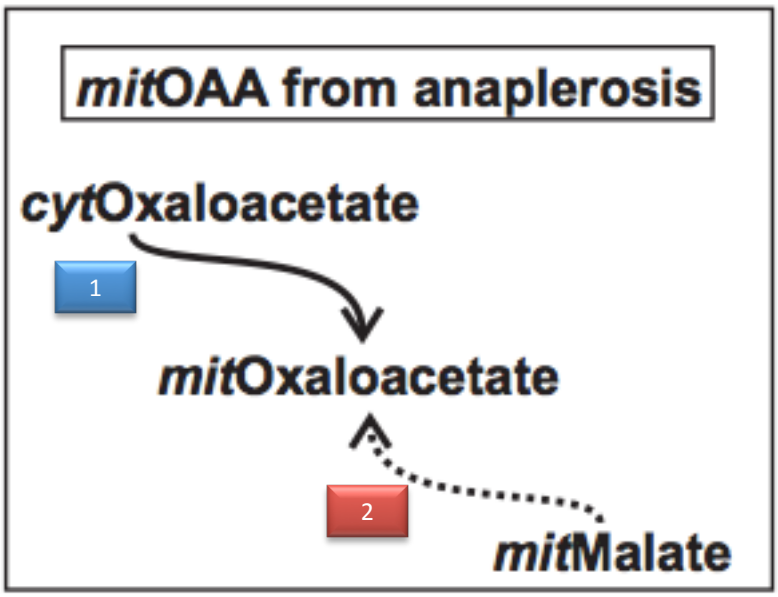

Figure 28 Illustration of how mitOAA from anaplerosis flux ratio was calculated.

Table 17 Summary of data used to calculate flux ratio 'mitOAA from anaplerosis'.

Pathway	Flux Formula	Reaction Formula
1: mitOAA from cytOAA	rxn1211	'H + Oxaloacetate <=> H + Oxaloacetate '
2: mitOAA from mitMalate	rxn1075	'L_Malate + Nicotinamide_adenine_dinucleotide <=> H + Nicotinamide_adenine_dinucleotide___reduced + Oxaloacetate '

5.3 *E. coli*

Haverkorn et al.[31] applied large-scale 13C-flux analysis to 91 transcriptional regulator mutants of *Escherichia coli* under two growth conditions: glucose minimal media and galactose minimal media. Such choice of external conditions was motivated by two distinct modes of hexose catabolism under those conditions. Figure 29 represents the pathway usage on both substrates.

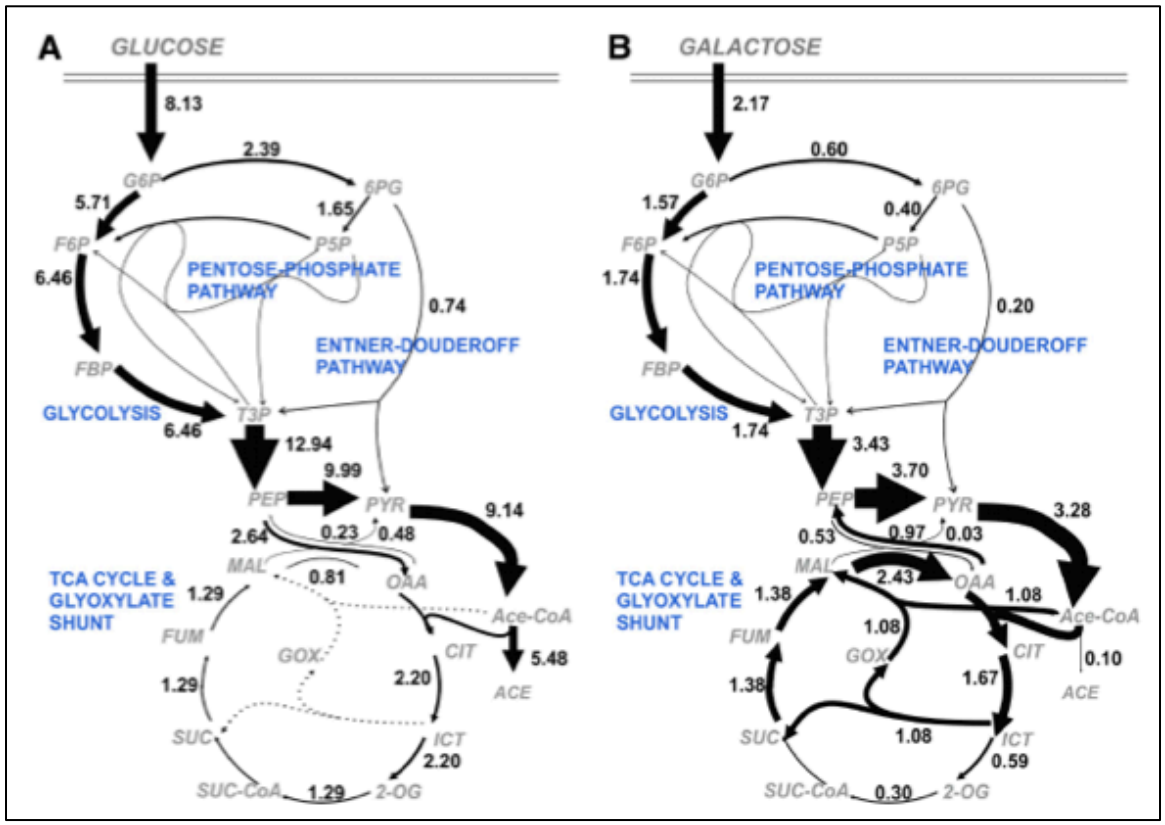

Figure 29 Representation of two distinct modes of hexose catabolism on glucose and galactose. Arrows represent absolute metabolic fluxes during aerobic growth.[31]

The estimated, via FiatFlux[35] software, absolute fluxes for each TF knockout varied greatly. However, in contrast to absolute fluxes, the relative flux partitioning was rather invariant in the entire mutant set[31]. The recorded significant changes in flux ratios occurred around the acetyl-CoA branch point. Thus, in this study we considered only a subset of all published flux ratio

studies focusing on those which should show the greatest changes in flux direction. The examined cases were: (1) oxaloacetate from PEP; (2) PEP from oxaloacetate; and (3) oxaloacetate from glyoxylate shunt. The analysis was performed in an analogous way to *E.coli* study. Tables 18 and 19 summarize the data used to simulate relative flux ratios. Results are presented in Table 20

Table 18 Summary of data used to calculate flux ratios for oxaloacetate originating from PEP; PEP originating from oxaloacetate; and oxaloacetate originating from the glyoxylate shunt.

Flux ratio (pathway 1/pathway 2)	Pathway	Flux Formula
OAA from PEP	**1**: OAA fom PEP **2**: OAA from Malate	**1**: rxn2181-rxn2183 **2**: rxn1758
PEP from OAA	**1**: PEP from OAA **2**: PEP from 3-P-Glycerate	**1**: rxn2183-rxn2181 **2**: rxn1102
OAA from GS	**1**: OAA fom isocitrate **2**: OAA from PEP	**1**: rxn1714+rxn2404 **2**: rxn2181-rxn2183

Table 19 Reaction formulas for all reactions used to calculate flux ratios in *E. coli* study.

Reaction Number in Metabolic model	Reaction Formula
1102	'D-Glycerate 2-phosphate <=> H2O + Phosphoenolpyruvate '
1714	'Acetyl-CoA + Glyoxylate + H2O -> Coenzyme A + H+ + L-Malate '
1758	'L-Malate + Nicotinamide adenine dinucleotide <=> H+ + Nicotinamide adenine dinucleotide - reduced + Oxaloacetate '
2181	'CO2 + H2O + Phosphoenolpyruvate -> H+ + Oxaloacetate + Phosphate '
2183	'ATP + Oxaloacetate -> ADP + CO2 + Phosphoenolpyruvate '
2404	'2-Oxoglutarate + O2 + Taurine -> Aminoacetaldehyde + CO2 + H+ + Sulfite + Succinate '

Table 20 Summary of the number of correctly predicted flux changes for each considered flux ratio obtained via PROM algorithm.

Flux ratio	Number of correctly predicted flux ratio changes on glucose (out of 80)	Number of correctly predicted flux ratio changes on galactose (out of 80)
OAA from PEP	21	4
PEP from OAA	41	12
OAA from GS	NA	1

5.4 Conclusions

In this chapter we explored the prospective application of PROM's algorithm to predict flux changes in addition to growth phenotypes. Two sets of experimentally determined flux ratios for *S. cerevisae* and *E. coli* were used as validation sets. Those values were quantified using well-established 13C-labeling method[33]; thus, experimental values were treated as high-confidence data. PROM predictions of flux ratio direction change for both cases had small overlap percentage with experimental values (about 30-60% for *S. cerevisae* and 1-50% for *E. coli*). Thus, PROM was determined not to be a feasible tool for flux ratio predictions. This poor performance could have been caused by small variability in reaction rates ratios. Thus, PROM's sensitivity was too low to detect those changes. In addition, Haverkorn et al.[31] realized that, in contrast to small variability in flux ratio changes, the absolute flux ratios vary greatly for TFs KOs and diverse growth conditions.

6. Assessment of the PROM's performance

In this paper we analyzed the ability of PROM's algorithm to generate high-confidence genome-scale metabolic-regulatory networks. Specifically, we discovered the impact of the quality and amount of required data (metabolic model, transcriptional regulatory network, and gene expression data), on the accuracy of the generated reconstruction of metabolic and regulatory interactions occurring within a cell. The available growth phenotype data for *E. coli*, *M. tuberculosis* and *S. cerevisae* was used to validate the computational model. A combination of gene essentiality data for optimal growth, gene lethality data, and experimentally measured growth rates for various transcriptional regulatory mutants under different conditions was used to assess PROM's performance. The first two studies resulted in measuring accuracy, sensitivity, specificity, and normalized score while the later one gave correlation between experimental and computationally predicted growth.

We showed that high quality metabolic and transcriptional regulatory networks reconstructions were crucial for PROM to be predictive. However, accurately represented gene-transcription factor (TF) interactions had more significant effect than metabolic model alone. Also, those interactions had to be determined experimentally and not through mutual information based inference algorithms such as ASTRIX. In case of gene expression data, the optimal number of microarray experiments was observed for each organism that resulted in no further improvement of PROM's performance.

This study revealed that PROM's algorithm could be successfully applied to model less complex bacterial systems (*E. coli* and *M. tuberculosis*) with growth correlations greater than 0.50, but when used to simulate eukaryotic cells (*S. cerevisae*) its accuracy is very poor (correlations in the range of 0.24-0.26). Also, the incorporation of flux variability analysis (FVA) to model flux changes after TF KOs proved that alternative methods should be explored since the percentage of correctly predicted changes varied greatly from 1-60% and was only slightly better than random for few cases.

Up to this point, only three genome-scale metabolic-regulatory reconstructions were performed: two for bacteria and one for a eukaryote. Thus, to further examine PROM algorithm's strengths and weaknesses in modeling metabolic-regulatory intracellular interactions, additional reconstructions need to be generated for a wide array of organisms. Moreover, other approaches to modify PROM for the prediction of reaction rates should be analyzed. Specifically, exploring absolute fluxes instead of flux ratios should be considered.

References

[1] Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2:343-72.

[2] Feist AM, Herrgard MJ, Thiele I, Reed JL, Palsson BO (2009) Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol* 7:129-143.

[3] Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. *Mol Syst Bio* 5:320.

[4] Herrgard MJ, Covert MW, Palsson BO (2004) Reconstruction of microbial transcriptional regulatory networks. *Curr Opin Biotechnol* 15:70-77.

[5] Chandrasekaran S, Price ND (2010) Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in *Escherichia coli* and *Mycobacterium tuberculosis*. *Proc Natl Acad Sci USA* 107(41):17845-50.

[6] Covert MW, Xiao N, Chen TJ, Karr JR (2008) Integrating metabolic, transcriptional regulatory and signaling transduction models in *Escherichia coli*. *Bioinformatics* 24(18):2044-2050.

[7] Covert MW, Palsson BO (2002) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J Biol Chem* 277(31): 28058-28064.

[8] Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? *Nat Biotechnol* 28: 245-248.

[9] Price ND, Papin JA, Schilling CH, Palsson BO (2003) Genome-scale microbial in silico models: The constraints-based approach. *Trends Biotechnol* 21:162-169.

[10] Herrgard MJ, Lee BS, Portnoy V, Palsson BO (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res 16*(5):627-635.

[11] Goelzr A, et al, (2008) Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. *BMC Syst Biol* 2:20.

[12] Chandrasekaran S, Ament SA, Eddy JA, Rodriguez-Zas SL, Schatz BR, Price ND, Robinson GE (2011) Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. *Proc Natl Acad Sci* USA 29.

[13] Mahadevan R, Schilling CH (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 5:264-276.

[14] Feist AM et al. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3:127.

[15] Jamshidi N, Palsson BO (2007) Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the *in silico* strain iNJ661 and proposing alternative drug targets. *BMC Systems Biology* 1:26.

[16] Salgado H, et al. (2004) RegulonDB (version 4.0): Transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res* 32(Database issue):D303-D306.

[17] Balazsi G, Heath AP, Shi L, Gennaro ML (2008) The temporal response of the *Mycobacterium tuberculosis* gene regulatory network during growth arrest. *Mol Syst Biol* 4:225.

[18] Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429:92-96.

[19] Sassetti CM, Boyd DH, Rubin EJ (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol Microbiol* 48:77-84.

[20] Gao Q et al. (2005) Gene expression diversity among Mycobacterium tuberculosis clinical isolates. *Microbiology* 151:5-14.

[21] Lamichhane G, et al. (2003) A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis. *Proc Natl Acad Sci* USA 100:7213-7218.

[22] Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, Palsson BO (2011) A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism – 2011. *Mol Syst Biol* 7:535.

[23] Gama-Castro S, et al. (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated with genetic sensory response units (Gensor Units). *Nucleic Acids Res* 39(Database issue):D98-D105.

[24] Fang X, Wallqvist A, Reifman J (2010) Development and analysis of *in vivo* – compatible metabolic network of *Mycobacterium tuberculosis*. *BMC Syst Biol* 4:160.

[25] Sanz J, Navarro J, Arbues A, Martin C, Marijuan PC, Moreno Y (2011) The transcriptional regulatory network of Mycobacterium tuberculosis*. PLoS One* 6(7):e22178.

[26] Griffin JE, Gawronski JD, Dejesus MA, Ioerger TR, Akerley BJ, Sassetti CM (2011) High-resolution phenotypic profiling defines genes essential for mycobacterial growth and cholesterol catabolism. *PLoS Pathog* 7(9):e1002251.

[27] Zomorrodi AR, Maranas CD (2010) Improving the iMM904 *S. cerevisiae* metabolic model using essential synthetic lethality data. *BMC Syst Biol* 4:178.

[28] Teixeira MC et al. (2006) The YEASTRACT database: a tool for the analysis of transcriptional regulatory associations in Saccharomyces cerevisaie. *Nucl Acids Res* 34:D446-D451.

[29] Margolin AA, et al. (2006) ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7(Suppl 1):S7.

[30] Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. *Ann Stat* 32:407-499.

[31] Haverkorn van Rijsewijk BRB, Nanchen A, Nallet S, Kleijn RJ, Sauer U (2011) Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in *Escherichia coli*. *Mol Syst Biol* 7:477.

[32] Fendt SM, Oliveira AP, Christen S, Picotti P, Dechant RC, Sauer U (2010) Unraveling condition-dependent networks of transcription factors that control metabolic pathway activity in yeast. *Mol Syst Biol* 6:432.

[33] Sauer U (2006) Metabolic networks in motion: 13C-based flux analysis. *Mol Syst Biol* 2:62.

[34] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Res* 40:D109-D114.

[35] Zamboni N, Fischer E, Sauer U (2005) FiatFlux – a software for metabolic flux analysis from 13C-glucose experiments. *BMC Bioinformatics* 6:2009.